Australian Bureau of Statistics

2080.5 - Information Paper: Australian Census Longitudinal Dataset, Methodology and Quality Assessment, 2006-2016

Latest ISSUE Released at 11:30 AM (CANBERRA TIME) 20/03/2019

Summary

Introduction

INTRODUCTION

The Australian Census Longitudinal Dataset (ACLD) uses data from the Census of Population and Housing to build a rich longitudinal picture of Australian society. The ACLD can uncover new insights into the dynamics and transitions that drive social and economic change over time, and how these vary for diverse population groups and geographies. Three waves of data have contributed to the ACLD so far, from the 2006, 2011 and 2016 Censuses.

There are two ACLD panels, representing a 5% sample of records from the 2006 Census and the 2011 Census, respectively. The 2006 Panel comprises of records from the original 2006 ACLD sample linked to records from the 2011 Census and the 2016 Census. The 2011 Panel is linked to records from the 2016 Census.

As new panels and information from subsequent Censuses are added to the ACLD, its value as a resource for longitudinal studies of the Australian population will continue increasing.

This paper describes the background and rationale for the ACLD, the data linkage methodology used for producing the 2006 and 2011 ACLD panels and an assessment of its quality.

1.1 OVERVIEW

Development

In 2005, the ABS embarked on a project to enhance the value of Census data by bringing it together with other datasets, both ABS and non-ABS, to leverage more information from the combination of datasets than would be available from the individual datasets separately. The ACLD was proposed as an enduring longitudinal dataset constructed through the linking of records from successive Censuses.

As part of the development phase, a quality study was undertaken in which data from the 2005 Census Dress rehearsal were linked to data from the 2006 Census. This quality study concluded that the linkage methodology was feasible and that the expected quality of the linked data file would be sufficient for longitudinal analysis. For more information see, Assessing the Likely Quality of the Statistical Longitudinal Census Dataset (cat. no. 1351.0.55.026).

In 2013 the ABS released the first ACLD product, a 5% sample of the 2006 Census linked to the 2011 Census (the 2006 ACLD Panel). In preparation for adding 2016 Census data to the ACLD, a new panel of 2011 Census records was selected as a representative sample of the 2011 Census population. The 2011 Panel was designed to include:

- most of the 2011 Census records that were linked in the 2006 Panel;

- new records to account for missed links in the 2006 Panel; and
- new records to represent new births and migrants since the 2006 Census.

The 2011 Panel size was increased slightly to 5.7%, to achieve a linked sample size of no greater than 5% of the population after allowing for missed links and people in the 2011 sample not being in scope of the 2016 Census due to death or overseas migration (note that the linked sample size for the 2006 Panel linked to the 2011 Census was only 4.2%.) The 2011 ACLD Panel was released in 2018, consisting of the 2011 Panel sample of records from the 2011 Census linked to the 2016 Census.

In the March 2019 release, the 2006 Panel has been re-linked to the 2011 Census to take advantage of improved linking methodology since the initial release, and has then been linked to records from the 2016 Census.

Linking the ACLD

Data linkage is typically undertaken using a combination of deterministic and probabilistic methods:

- Deterministic linkage involves assigning record pairs across two datasets that match exactly or closely on common variables. This type of linkage is most applicable where the records from different sources consistently report sufficient information and can be an efficient process for conducting linkage
- Probabilistic linkage is based on the level of overall agreement on a set of variables common to the two datasets. This approach allows links to be assigned in spite of missing or inconsistent information, providing there is enough agreement on other variables.

For many individuals the linkage process will have accurately matched their corresponding records between Censuses. In some cases, the link will represent different people who share a number of characteristics in common. Some inaccuracy in the linkage will not generally affect statistical conclusions drawn from the linked data, although care should be taken in the interpretation of results. For more information see Section 2 - Data Linking Methodology.

1.2 MULTI-PANEL SAMPLE DESIGN

Without sample maintenance, the ACLD would decline in its ability to accurately reflect the Australian population over time due to:

- people newly in scope of the ACLD (i.e. children born and immigrants arrived in Australia since the previous Census) not being represented in the sample;
- people selected in the ACLD sample no longer being in scope due to death or overseas migration; and
- missing and/or incorrect links (linkage bias).

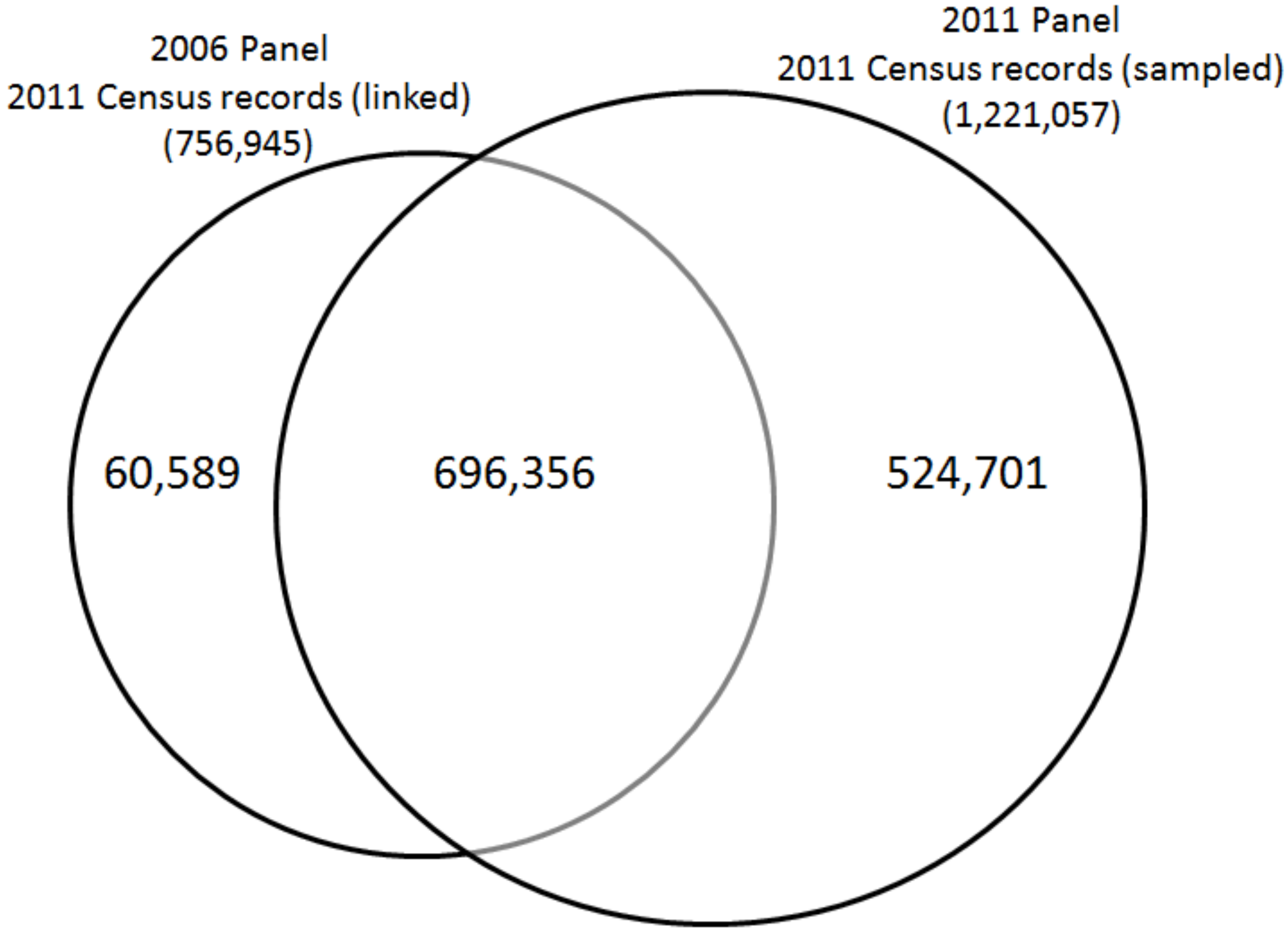
Linkage bias in longitudinal datasets is unique to those created via data integration, as traditional longitudinal studies employ strategies to ensure they collect information about the same individual over time. In a linked longitudinal dataset, data integration is necessary due to a lack of a common identifier to identify a person's responses over time. Linkage bias occurs where certain populations are more difficult to link than others (e.g. Aboriginal and Torres Strait Islander people, young males), so links are more likely to not be identified for members of these groups and, if they are found, have a higher chance of being inaccurate. If left untreated, the representation of population groups suffering from linkage bias would worsen as each new Census is linked to the ACLD.

The ACLD sample is maintained through application of the Multi-Panel framework, developed by Chipperfield, Brown & Watson (2017). This framework provides an approach for selecting records in the ACLD to create panels which maintain the longitudinal and cross-sectional representativeness of the dataset over time, while minimising the impact of accumulated linkage bias on longitudinal analysis.

The Multi-Panel approach designs multiple overlapping panels, with each panel representing a single Census population (2006, 2011, 2016, etc.), which is then linked to subsequent Censuses. The sample selection strategy for each panel is designed to maintain a linked sample size of 5%, maximise sample overlap between the panels, and introduce new records to the dataset in each panel to account for new births, migrants and missed links in previous panels. This allows flexibility for users, who can draw on the most appropriate panel for their research question.

The sample overlap between the 2006 and 2011 ACLD Panels is illustrated below:

FIGURE 1 - SAMPLE OVERLAP BETWEEN THE 2006 AND 2011 ACLD PANELS



1.3 ACCESS TO THE ACLD

The ACLD is accessible online through ABS TableBuilder and DataLab. Through ABS TableBuilder clients can build, customise, save and export their own tables and graphs. In this product, confidentiality methods are applied to the data prior to output to ensure that information that is likely to enable identification of an individual or household will not be released. The DataLab is an interactive data analysis solution available for high end users to run advanced multivariate statistical analyses, for example, multiple regressions and structural equation modelling. The DataLab environment contains up to date versions of SPSS, Stata, SAS and R analytical languages. Controls in the DataLab have been put in place to protect the identification of individuals and organisations. These controls include environmental protections, data de-identification and confidentialisation, access safe guards and output clearance. All output from DataLab sessions is cleared by an ABS officer before it is released.

For more information, or to access the ACLD, see Microdata: Australian Census Longitudinal

Dataset, ACLD (cat. no. 2080.0).

Data Linking Methodology

2. DATA LINKING METHODOLOGY

The data linking process used to create the ACLD included a series of steps which can be generalised into the following:

- standardisation of data;
- preparation of data;
- blocking;
- record pair comparison; and
- decision model.

2.1 STANDARISATION OF DATA

Before records on two datasets are compared, the contents of each need to be as consistent as possible to facilitate comparison. This process is known as 'standardisation' and includes a number of steps including verification, recoding and re-formatting variables, and parsing text variables (i.e. separating text variables into their components). Additionally, some variables such as name may require substantial repair prior to standardisation.

Some variables, such as age, differ between the two datasets in a predictable way, and an adjustment is required to account for this variance. Some variables are coded differently at different points in time, and concordances may be necessary to create variables which align on the two datasets. Variables may also be recoded or aggregated in order to obtain a more robust form of the variable. Standardisation takes place in conjunction with a broader evaluation of the dataset, in which potential linking variables are identified.

The standardisation procedure for the ACLD linkage project involved coding imputed and invalid values for selected variables to a missing value. These variables include name, address, day of birth, month of birth, year of birth, age, sex, year of arrival and marital status. Standardisation for hierarchical variables involved collapsing higher levels of aggregation to minimise disagreement when linking records which may have had a small intercensal change or to account for potential differences in the coding of the variable. This allows for records to agree using broader categories rather than disagree on specific information that may have changed over time or be reported and/or coded inconsistently. An example of this is country of birth. Whereas in 2011 the respondent may have been coded to 'Northern Europe' (two digit level of country of birth), in 2016 they may have reported a specific country such as 'England' or 'Norway' (four digit level of country of birth). If left in its original state, a comparison between 'Northern Europe' and 'England' would not agree, even though one is a sub-category of the other. Both two digit and four digit versions of country of birth were used in the linkage, albeit in different passes. This improved the quality of the linkage while also increasing the chances that a link would be made.

The following is a description of specific standardisation techniques that were performed on variables for this project.

First Name and Surname (applies to the 2011 and 2016 Censuses)

In the 2011 Census, name data was first subjected to an automated repair process. Both first names and surnames were compared against corresponding master name indexes, with names being repaired when a suitably close match to a value on the index was found. The name repair process was repeated for the 2016 Census data, with the addition of a number of enhancements.

These enhancements optimised both the number and accuracy of names repaired, and included the following:

- expanded first name index, enabling a larger amount of names to be repaired;
- identification and removal of values that were not considered to be a valid name;
- use of age and country of birth information to assist with repairing first names;
- use of family structure to assist with repairing surnames;
- targeted automatic repair processes based on response type (i.e. online forms vs. paper forms);
- a manual coding process for paper form responses that could not be sufficiently repaired through automatic means (note that manual repair was performed only for a subset of records in 2011); and
- retention of additional repaired name options when more than one close match for a name value was found on the index.

After repair, first names were then compared against a nickname concordance, ensuring that different variations would be grouped into a common name for the purposes of linkage. The standardisation of the same name value may also vary depending on the reported sex. For example, a female with the name 'Jess' may be standardised to 'Jessica' whereas it may be standardised to 'Jesse' for a male. Any first names that could not be matched to a nickname retained their original form. An identical nickname concordance was used in both the 2011 and 2016 Censuses, ensuring that the name values from both Censuses were consistently standardised. The nickname concordance process was not performed for surnames.

After Census names had been repaired and standardised, they were converted into anonymous hash codes to be used in the linking process. These encoded versions of Census name served to assist in further protecting the privacy of Census respondents when linking 2011 and 2016 Census records in the ACLD. No name information was retained from the 2006 Census, so no name information was used in linking 2006 and 2011 Census records.

The hash codes are created by grouping people with a combination of letters from their first and last names using a secure one-way process, meaning that a code cannot be reversed to deduce the original name information. Each code represents approximately 2,000 people drawn from many different letter combinations, and therefore is not unique to an individual. Encoding of 2011 and 2016 Census name information was undertaken during the processing of each respective Census. The encoded name information is retained by the ABS for linking purposes. Original Census name information was not used in linking the ACLD.

The codes are only accessible to those ABS officers creating the linked dataset, and will never be released outside the ABS.

Geography

As a proportion of the Australian population is expected to change their residential address between Censuses, the ACLD uses geographic data expected to refer to the same time point in the linking process. Usual address from one Census is compared to usual address five years ago from the subsequent Census. Additionally, the following standardisation techniques are applied:

- imputed address geography is removed for linking purposes, for example a mesh block which was imputed will not be used for linking, though it will remain on the analytical file;
- usual address from a particular Census was converted to the Australian Statistical Geography Standard (ASGS) that is relevant for the subsequent Census to allow for accurate comparison of Census address data. For example, 2011 Census address was converted to the 2016 ASGS;
- where a 2006 ASGS unit had been split into multiple 2011 ASGS units, only the 2011 ASGS unit with the most overlap was retained for linkage, while in instances where a 2011 ASGS unit had been split into multiple 2016 ASGS geographic units, up to three options were retained for linkage. For further information on changes to ASGS geography refer to

Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, July 2011 (cat. no. 1270.0.55.001) for 2011 ASGS or Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, July 2016 (cat. no. 1270.0.55.001) for 2016 ASGS; and,

- to increase the chance of linking on geography and minimise the impact of respondent recall error concerning address five years ago, where a 2011 or 2016 Census respondent reported a different usual address or address one year ago to address five years ago, these addresses were used as alternative options in the linking process when linking to the previous Census.

Personal Characteristics

- Invalid dates of birth were removed.
- Imputed instances of sex, marital status and age were removed.
- Age increased by 5 years on the initial Census involved in the respective linkage. For example, when linking the 2006 Panel to the 2011 Census, records from 2006 had their age increased by 5. The same was done for linkage from 2011 to 2016 records.
- Country of birth coded to the two digit level, for example 'Western Europe', to improve chances of linkage. Four digit country of birth, for example 'Austria' and 'Germany' was also retained to increase quality of links that agreed on this level of geography. Each level of country of birth was used in different passes.
- Indigenous status standardised to group 'Aboriginal', 'Torres Strait Islander' and 'Both Aboriginal and Torres Strait Islander' as one unique response.
- Marital status standardised to group 'Divorced', 'Separated' and 'Widowed' as one unique response. In addition, marital status was coded to missing for persons under 15 years of age.

2.2 PREPARATION OF DATA

An additional data preparation technique was used for Census records where multiple responses had been provided for key linking variables. A record may have had multiple responses for a single linking variable in the following situations:

- a name that required repair returned more than one possible repaired name value from the applied process (2016 Census records only);
- the respondent reported different locations for address of usual residence, address one year ago and address five years ago (only for records in the most recent Census involved in a particular linkage, e.g. for 2016 Census records when linking to 2011 Census records); or,
- the process of aligning 2011 Census usual address geographies to 2016 ASGS values resulted in more than one possible set of these values for some records.

The process for allowing the use of multiple responses for a linking variable involved restructuring the data for affected records; multiple rows were created, with the number of rows generated equal to the number of different combinations that could be created from the linkage information. This is demonstrated in Tables 1a and 1b below. A respondent with two different encoded first name values and two different mesh blocks would have four permuted rows generated. Meanwhile, the information that only has one stated value (in this example surname and date of birth) is duplicated across all of the generated rows. Structuring the data in this manner allows for all combinations of a respondent's linkage information to be considered in a highly efficient manner. Permutation of name was only used for linkages of 2016 Census records to 2011 Census records as name data was not retained for the 2006 Census. Permutation of data was not used for the original 2006 Panel linkage.

TABLE 1A - EXAMPLE OF DATA RESTRUCTURE, Original Record

Person ID	Encoded First	Encoded First	Encoded	Encoded	Encoded Mesh	Encoded Mesh	Date of

	Name 1	Name 2	Surname 1	Surname 2	Block 1	Block 2	Birth
1	1234	5678	9876		12345670000	98765430000	09/08/2016

TABLE 1B - EXAMPLE OF DATA RESTRUCTURE, Restructured Record

Person ID	Encoded First Name	Encoded Surname	Mesh Block	Date of birth
1	1234	9876	12345670000	09/08/2016
1	1234	9876	98765430000	09/08/2016
1	5678	9876	12345670000	09/08/2016
1	5678	9876	98765430000	09/08/2016

2.3 RECORD PAIR COMPARISON

There were two different linking methods utilised in the linkage of the ACLD, deterministic and probabilistic. Deterministic linkage methods were initially used to identify matches that had high quality linking information. Probabilistic linking was then used in subsequent passes for the records that had not been linked in the deterministic passes. Probabilistic passes were linked non-sequentially (this method is explained further in Section 2.3.2 Probabilistic Linking).

2.3.1 Deterministic Linking

Deterministic data linkage, also known as rule-based linkage, involves assigning record pairs across two datasets that match exactly or closely on common variables. This type of linkage is most applicable where the records from different sources consistently report sufficient information to efficiently identify links. It is less applicable in instances where there are problems with data quality or where there are limited characteristics.

Initially, a deterministic linkage method was used to identify matches that contained high quality linking information. This involved using selected personal and demographic characteristics (first name hash code, surname hash code, sex, date of birth/age, mesh block and country of birth), to identify the highest quality record pairs that matched exactly on these characteristics. These links were accepted and exempted from the following probabilistic linkage passes. This method identified approximately 75% of links for each of the 2006-2011 Census and 2011-2016 Census linkages with the remaining 25% identified via probabilistic linking.

2.3.2 Probabilistic Linking

Probabilistic linking allows links to be assigned in spite of missing or inconsistent information, providing there is enough agreement on other variables to offset any disagreement. In probabilistic data linkage, records from two datasets are compared and brought together using several variables common to each dataset (Fellegi & Sunter, 1969).

A key feature of the methodology is the ability to handle a variety of linking variables and record comparison methods to produce a single numerical measure of how well two particular records match, referred to as the 'linkage weight'. This allows ranking of all possible links and optimal assignment of the link or non-link status (Solon and Bishop, 2009).

Blocking variables

In probabilistic linkage, record pairs (consisting of one record from each file) can be compared to see whether they are likely to be a match, i.e. belong to the same person. However, if the files are even moderately large, comparing every record on File A with every record on File B is computationally infeasible. Blocking reduces the number of comparisons by only comparing record pairs where matches are likely to be found – namely, records which agree on a set of blocking variables. Blocking variables are selected based on their reliability and discriminatory power. For instance, sex is partially useful as it is typically well reported, however it is minimally informative as it only divides datasets into two blocks, and therefore does not sufficiently reduce the computational intensity of larger linkages. Accordingly, it is generally not used alone but rather in conjunction with other variables.

Comparing only records that agree on one particular set of blocking variables means a record will not be compared with its match if it has missing, invalid or legitimately different information on a blocking variable. To mitigate this, the linking process is repeated a number of times ('passes'), using a range of different blocking strategies. For example, on the first probabilistic pass of the 2011-2016 linking strategy, a block using a fine level of geography (mesh block) was used to capture the majority of 2011 Census records that had matching information with their corresponding 2016 Census record. The second probabilistic pass blocked on a slightly broader level of geography (SA1), to capture records which disagreed on mesh block, but had matching information at the higher geographic level. The blocking variables used for each pass are outlined in Section 2.4 Blocking and Linking Strategy used in the ACLD.

Linking variables

Within a blocking pass, records on the two files which agree on the specified blocking variables are compared on a set of linking variables. Each linking variable has associated field weights, which are calculated prior to comparison. Field weights indicate the amount of information (agreement, disagreement, or missing values) a linking variable provides about whether or not the records belong to the same person (match status). Field weights are based on two probabilities associated with each linking variable: first, the probability that the field values agree given that the two records belong to the same person (match); and second, the probability that the field values agree given the two records belong to different persons (non-match). These are called *m* and *u* probabilities (or match and non-match probabilities) and are defined as:

$$m = \text{P(fields agree | records belong to the same person)}$$
$$u = \text{P(fields agree | records belong to different people)}$$

Given that the *m* and *u* probabilities require knowledge of the true match status of record pairs, they cannot be known exactly, but rather must be estimated. The ABS calculated the *m* and *u* probabilities based on a training dataset, under the assumption that each deterministic link on the dataset was a match. The deterministic links used in this phase included (1) the highest quality links accepted in the deterministic linking passes, and (2) additional slightly lower quality links expected to be confirmed in the probabilistic linking phase. This method estimated the likelihood that a record would have a match by taking deaths and net overseas migration into account when estimating the *m* and *u* probabilities. This method also generated probabilities for disagreement, which can be referred to as *md* and *ud* probabilities:

$$md = \text{P(fields disagree | records belong to the same person)}$$
$$ud = \text{P(fields disagree | records belong to different people)}$$

Note that *m* and *u* probabilities are calculated separately for each pass, as the probabilities depend upon the characteristics of the pass' blocking variables. For example, the *m* probability for country of birth when blocking on mesh block will be different to the *m* probability for country of birth when blocking on sex.

Match (*m*) and non-match (*u*) probabilities are then converted to agreement and disagreement field weights. They are as follows:

$$\text{Agree} = \log2(m/u)$$
$$\text{Disagree} = \log2(md/ud)$$

These equations give rise to a number of intuitive properties of the Fellegi–Sunter framework (Fellegi & Sunter, 1969). First, in practice, agreement weights are always positive and disagreement weights are always negative. Second, the magnitude of the agreement weight is driven primarily by the likelihood of chance agreement. That is, a low probability of two random people agreeing on a variable (for example, date of birth) will result in a large agreement weight being applied when two records do agree.

The magnitude of the disagreement weight is driven by the stability and reliability of a variable. That is, if a variable is well reported and stable over time (for example, sex) then disagreement on the variable will yield a large negative weight. For each record pair comparison, the field weights from each linking variable are summed to form an overall record pair comparison weight or 'linkage weight'.

Before calculating *m* and *u* probabilities for some variables it is first necessary to define what constitutes agreement. Typical comparison functions used in the ACLD linkage include:

- Exact match (e.g. Sex). Agreement occurs only when the two variable values are identical. This criterion is used for most linking variables; and
- Numeric difference (e.g. Age). A pair may be defined to agree if their variable values differ by an amount less than or equal to a specified maximum difference.

For further details on comparison functions used for probabilistic linkage, see Christen & Churches (2005).

Near or partial agreement may also be factored into the linking process through calculation of *m* and *u* probabilities accounting for such agreement. For example, a person’s age on equivalent records will frequently be an exact match, and the *m* and *u* probabilities are calculated based on this definition. During linkage, however, a partial agreement weight was given for age within one year difference to cater for persons who may have understated their age in one Census and/or overstated it in the following Census or vice versa.

Blocking variables, linking variables, comparator types, and *m* and *u* probabilities are used as input parameters for the linking software. Records which agree on the blocking variable(s) are compared on all linking variables.

2.4 BLOCKING AND LINKING STRATEGY USED IN THE ACLD

The strategy employed for the re-link of the 2006 Panel to the 2011 Census and linking of both ACLD panels to the 2016 Census builds on the original 2006-2011 ACLD linking strategy, using developments in linking methodology, software and available data to improve the approach. For further details on the original 2006-2011 ACLD linkage see Linkage Results.

To develop the linking strategy to be used for the 2011-2016 ACLD linkage, the 2006 Panel of the ACLD was re-linked to the 2011 Census as part of an investigation into the feasibility of proposed methodological enhancements. While the re-link of the 2006 ACLD Panel sample could not make use of hash encoded name (as the 2011-2016 linkage would benefit from), it was found that improvements could be made on the original linkage with regards to the estimated precision and accuracy of the links achieved. The enhanced linking strategy was then implemented for both the 2006-2011 and 2011-2016 linkages.

The key features of the enhanced linking strategy used include:

- a combination of deterministic and probabilistic linkage techniques designed to link a high quality dataset that is representative of the Australian population;
- linking variables found to contain unacceptably low levels of consistent reporting over time, such as highest year of schooling and occupation, were removed from the linking strategy;
- certain passes were designed to link particular population groups in order to improve linkage rates, such as Aboriginal and Torres Strait Islander peoples, migrants, and children;
- use of a non-sequential approach to probabilistic linking. The sequential approach used for

the original 2006-2011 ACLD linkage removed accepted links after each probabilistic pass, resulting in the successful identification of true matches being dependent on the order of the passes. The non-sequential approach allows for all records to be given an opportunity to link in every probabilistic pass, preventing poorer quality links from earlier passes being accepted where a higher quality link could be found in a later pass; and,

- blocking weights were applied to each of the probabilistic passes to standardise the linkage weights for all potential record pairs. This allowed all potential links to be comparable across passes to determine the best possible link for each record.

Table 2a displays the original linking strategy used for the 2006-2011 linkage for reference. Tables 2b and 2c display the blocking and linking variables applied in the 2006-2011 (re-link) and 2011-2016 linkages for each pass.

TABLE 2a - BLOCKING AND LINKING VARIABLES, By Pass Number, 2006 Panel, 2006-2011 linkage (original)

PASS NUMBER (a)(b)(c)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10(d)**	**1112**	
PERSONAL INFORMATION												
Age	B	B	L	L	L	L	L	L	L	L	L	L
Sex	B	B	B	B	B	B	B	B	L	L	L	L
Day and Month of Birth	B	B	B	B	B		L	L	L	L	L	L
Indigenous status			B	B	B	L	B		L	L	L	L
Country of Birth			L	L	L	L			L	L	L	
Year of Arrival			L	L	L				L	L	L	
Marital status			L	L	L		L	L				
Level of Qualification			L	L	L				L	L		
Field of Qualification			L	L	L		L	L	L	L	L	
Highest Level of Schooling			L	L	L				L	L		
Occupation							L	L				
Language spoke at home			L	L	L				L	L		
Religion		B	L	L	L							
Aged less than 15 block	B	B				B						
HOUSEHOLD INFORMATION												
Mother's Age		B				L						L
Mother's Day and Month of Birth		B				L						L
Father's Age												L
Father's Day and Month of Birth												L
Family ID block											B	
GEOGRAPHICAL INFORMATION												
Mesh Block	B		B				B			B		
SA1								B	B			
SA2		B		B								
SA4					B	B						B

(a) Passes 1 and 2 refer to the deterministic linking passes while passes 3-12 refer to the probabilistic linkage passes.
(b) B – blocking variable
(c) L – linking variable
(d) The results of Pass 10 were used to identify the blocking field to be used in Pass 11. As a result, there were no records output from Pass 10.

TABLE 2b - BLOCKING AND LINKING VARIABLES, By Pass Number, 2006 Panel, 2006-2011 linkage (re-link)

PASS NUMBER (a)(b)(c)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
PERSONAL INFORMATION								
Age	B	+/- 1	L	L			L	
Sex	B	L	L	L	B		B	B
Day and Month of Birth	B	L	L		B	B		B
Year of Birth	B			B	B	B		B
Indigenous status	B	L	L	B		L	L	
Country of Birth	B			B				
Year of Arrival		L	L	L			L	L
Marital status								
Language spoke at home				L			L	L
Religion							L	L
Aged less than 15 block					B			
HOUSEHOLD INFORMATION								
Mother's Age				L				
Mother's Day and Month of Birth					L			
Mother's Country of Birth				L	L			
Father's Age				L				
Father's Day and Month of Birth					L			
Father's Country of Birth				L	L			
GEOGRAPHICAL INFORMATION								
Mesh Block	B	B						
SA1			B					
SA2							B	
SA4					B	B		B

(a) Pass 1 refers to the deterministic linkage passes while passes 2-8 refer to the probabilistic linkage passes.
(b) B – blocking variable
(c) L – linking variable

TABLE 2c - BLOCKING AND LINKING VARIABLES, By Pass Number, 2006 and 2011 Panels, 2011-2016 linkage

PASS NUMBER (a)(b)(c)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

PERSONAL INFORMATION									
First name hash	B	L	L	L	L	L	L	L	L
Surname hash	B	L	L	L	L	L	L	L	L
Age	B	+/- 1	L		L			L	
Sex	B	L	L	B	B	B		B	B
Day and Month of Birth	B	L	L	B		B	B		B
Year of Birth	B			B		B	B		B
Indigenous status	B	L	L	L			L	L	
Country of Birth	B								
Year of Arrival		L	L	L				L	L
Marital status									
Language spoke at home				L				L	L
Religion				L				L	L
Aged less than 15 block					B	B			
HOUSEHOLD INFORMATION									
Mother's Age					L				
Mother's Day and Month of Birth						L			
Mother's Country of Birth					L	L			
Father's Age					L				
Father's Day and Month of Birth						L			
Father's Country of Birth					L	L			
GEOGRAPHIC INFORMATION									
Mesh Block	B	B							
SA1			B						
SA2					B			B	
SA4						B	B		B

(a) Pass 1 refers to the deterministic linkage passes while passes 2-9 refer to the probabilistic linkage passes.
(b) B – blocking variable
(c) L – linking variable

2.5 DECISION MODEL

In deterministic linking, an exact match is required on each of the variables specified in the blocking and linking strategy (see Passes 1 and 2 in Table 2a, and Pass 1 in Tables 2b and 2c). Using this approach, links were only accepted where a unique single record pair was identified. Where a record was included in more than one possible pair, it was returned to the pool of unlinked records for subsequent probabilistic passes.

In probabilistic linking, once potential record pairs are generated and weighted, a decision algorithm determines whether the record pair is linked, not linked or should be considered further as a possible link. The generation of potential record pairs from probabilistic linking can result in the records on one dataset linking to multiple records on the other, resulting in a file of 'many-to-many' potential links. The first phase of the decision process involves assigning a record to its best possible pairing. This process is known as one-to-one assignment. Ideally (and often true in practice) each record has a single, unique best pairing, which is its true match.

In the past, ABS probabilistic linking projects (including the original 2006 Panel linkage) have typically used an auction algorithm to assign links optimally from the pool of all possible links. The auction algorithm maximises the sum of all the record pair comparison weights through alternative assignment choices, such that if a record A1 on File A links well to records B1 and B2 on File B, but record A2 links well to B2 only, the auction algorithm will assign A1 to B1 and A2 to B2, to maximise the overall comparison weights for all record pairs.

For the 2016 ACLD linkage, a change was made to the assignment algorithm. Using the previous example, A1 may still link to B1, but A2 would only be able to link to B2 if it was a better link than A1 to B2. This change ensured that links would only be assigned when they are the absolute best option for both records in the link, which subsequently improved the quality of the links output at this phase. The modified algorithm was also far more efficient than the auction method, with the assignment process completed in a matter of minutes compared to several hours or days when using the auction algorithm. An additional change made for this ACLD linkage was that the one-to-one assignment was run using the combined many-to-many results from all passes in the linkage, rather than running the assignment over the results from each individual pass. This allowed the best links from all passes to be obtained from a single assignment procedure.

The second phase of the probabilistic decision rule stage takes the output of one-to-one assignment and decides which pairs should be retained as links, and which pairs should be rejected as non-links. The simplest decision rule uses a single 'cut-off' point, where all record pairs with a linkage weight above or at the cut-off are assigned as links, and all those pairs with a linkage weight below the cut-off are assigned as non-links. The best approach to assigning a single cut-off point is to clerically review links; however this process is time and resource intensive.

Model-based method

As clerical review was unavailable for the re-link of the 2006 panel to the 2011 Census due to data availability limitations, an alternative method of measuring precision and setting a cut-off was undertaken through the use of models. The method of Chipperfield et al (2018) was applied to provide an independent model-based estimate of the precision. The expected performance of this method was investigated using the 2011-2016 linkage and the results of clerical review undertaken for that linkage. While the clerical estimate of cumulative precision for the 2011-2016 linkage was 98.6%, the model-based approach estimated the precision to be over 99.0%. These results showed that the use of models was a viable option to generate a comparable estimate of precision where clerical review was not available. This model was used as the primary method of calculating precision and setting a cut-off for the 2006-2011 re-link for the 2006 Panel sample. Due to the lack of name information for the 2006 Census, the ability to distinguish a unique link became more difficult. To ensure a high quality linkage while maintaining a high linkage rate, the desired estimated cumulative precision was set at 95%, or an estimated false link rate of approximately 5%. This method achieved a 77.2% linkage rate when linking the 2006 Panel to 2011 Census records.

Clerical Review Method

In order to establish the cut-off point for the original 2006-2011 linkage and the 2011-2016 linkages, a sample of the record pairs were clerically reviewed. This provided the opportunity to ascertain quality levels and enabled an estimate of the number of 'false links', which are links formed that are believed to belong to separate entities (i.e. persons) rather than the same entity.

For the ACLD project, a sample of record pairs was clerically reviewed to set a single cut-off for the set of one-to-one links. Each sampled record pair was manually inspected to resolve its match status (i.e. if the link was 'true' or 'false'). As part of this process, a clerical reviewer was often able to use information which cannot be captured in the automated comparison process, but could be identified by the reviewer, such as common transcription errors (e.g. 1 and 7) or transposed information, such as the day of birth reported as the month or vice versa. This information was only available for the 2011 Census when conducting the original 2006 Panel linkage and for the 2016 Census when conducting the 2011-2016 linkages.

In addition to the linking variables, supplementary information was also used to confirm a link as true. This included:

- non-linking variables such as ancestry, occupation, schooling and qualification; and,
- reviewing the dates of birth and country of birth of parents (when available) for child records that had been linked.

These supplementary variables helped to clarify difficult decisions, especially on record pairs belonging to children, allowing for greater insight into whether a record pair was an actual match or just contained similar demographic and personal characteristics for two different individuals.

After completing the sample review, the results were used to set a single cut-off point for the 2011 Census to 2016 Census linkages, designed to assign a high proportion of links with high level of quality to the final linked dataset. This method achieved final 2011-2016 linkage rates of 80% for the 2006 Panel and 76% for the 2011 Panel.

ACLD 2006-11-16

PRODUCT OVERVIEW 2006-11-16

The 2006-11-16 ACLD is a representative sample of almost one million records from the 2006 Census (Wave 1) brought together with corresponding records from the 2011 Census (Wave 2) and the 2016 Census (Wave 3).

The 2006 Panel sample of records was originally linked to the 2011 Census and released in 2013. In this release, the 2006 Panel has been re-linked to the 2011 Census to take advantage of improved linking methodology since the initial release, and has then been linked to records from the 2016 Census.

The 2006-11-16 ACLD product is recommended for analysis of the 2006-11 and 2006-11-16 longitudinal populations.

Linkage Results

3. LINKAGE RESULTS, 2006-11-16, 2006 PANEL

At the completion of the linkage process 756,945 (77%) of the 979,662 records from the 2006 ACLD Panel sample were linked to a 2011 Census record to create the linked 2006-2011 ACLD file with an estimated precision of approximately 95%, or a false link rate of approximately 5%.

These record pairs were then linked to the 2016 Census via the 2011 Census record in each pair (any 2006 Panel record which had not been successfully linked to a 2011 Census record was not given the opportunity to link to the 2016 Census). This achieved 605,618 links (80% of the 2011 records in the 2006 Panel) at an estimated 98.6% precision for a direct 2011-2016 linkage. 62% of links from the original 2006 Panel sample linked to both the 2011 and 2016 Censuses.

All results presented in this section (unless identified in the relevant table) are based on characteristics from the 2006 ACLD Panel sample and have been confidentialised to prevent the identification of individuals.

Table 1 displays the linkage rate for a range of sub-populations.

TABLE 1 - LINKAGE RATES, By Selected Characteristics

	2006 Panel	**2006-11 Linked**	**2006-11**	**2006-11-16**	**2006-11-16**
	sample	**records**	**Linkage rate**	**Linked records**	**Linkage rate**
	(no.)	(no.)	(%)	(no.)	(%)
SEX					
Male	480 289	364 727	75.9	288 894	60.2
Female	499 376	392 222	78.5	316 724	63.4
AGE GROUP					
0-14	194 016	141 559	73.0	114 540	59.0
15-19	66 246	48 707	73.5	33 783	51.0
20-24	66 509	45 593	68.6	32 891	49.5
25-29	62 249	46 279	74.3	36 613	58.8
30-39	140 273	113 887	81.2	95 332	68.0
40-49	142 911	120 932	84.6	103 464	72.4
50-59	126 287	107 379	85.0	92 004	72.9
60-69	86 385	72 041	83.4	60 185	69.7
70-74	31 003	24 253	78.2	18 231	58.8
75 and over	63 781	36 305	56.9	18 583	29.1
INDIGENOUS STATUS					
Non-Indigenous	942 253	733 032	77.8	588 535	62.5
Aboriginal	19 697	12 449	63.2	8 765	44.5
Torres Strait Islander	1 451	940	64.8	674	46.5
Both Aboriginal and Torres Strait Islander	838	503	60.0	361	43.1
Not stated	15 421	10 020	65.0	7 283	47.2
STATE/TERRITORY OF USUAL RESIDENCE					
New South Wales	323 135	250 070	77.4	199 416	61.7
Victoria	244 097	191 981	78.6	154 283	63.2
Queensland	192 611	144 427	75.0	114 859	59.6
South Australia	75 476	59 386	78.7	48 422	64.2
Western Australia	95 795	73 948	77.2	59 261	61.9
Tasmania	23 781	18 624	78.3	14 815	62.3
Northern Territory	8 464	5 573	65.8	4 057	47.9
Australian Capital Territory	16 188	12 866	79.5	10 453	64.6
REMOTE AREAS					
Major Cities	669 274	523 474	78.2	420 691	62.9
Inner Regional	195 385	150 718	77.1	120 316	61.6
Outer Regional	92 397	68 940	74.6	54 283	58.7
Remote	13 988	9 844	70.4	7 537	53.9
Very Remote	6 548	3 905	59.6	2731	41.7
No Usual Address	2 024	0	0.0	0	0.0
Total(a)(b)(c)	**979 662**	**756 945**	**77.2**	**605 618**	**61.8**

(a) Data presented in the table have been perturbed. As a result, the sum of individual categories may not align with totals.
(b) Includes Other Territories.
(c) Includes Migratory areas.

The linkage rates for the 2006 ACLD panel were relatively consistent across most sub-populations and were in line with expected results. Compared with the overall linkage rate of 76%, the sub-populations which achieved the highest linkage rates for the 2006-2011 linkage were persons:

- aged 50 to 59 and 40 to 49 years (85%) and 60 to 69 years (83%);
- of non-Indigenous origin (78%);
- who usually lived in the Australian Capital Territory (80%); and
- who usually lived in major cities (78%) and inner regional areas (77%).

The same sub-populations had the highest linkage rates when linking through the three Census periods:

- aged 50 to 59 years (73%), 40 to 49 years (72%) and 60 to 69 years (70%);
- of non-Indigenous origin (63%);
- who usually lived in the Australian Capital Territory (65%); and
- who usually lived in major cities (63%) and inner regional areas (62%).

The sub-populations which achieved the lowest 2006-2011 linkage rates were persons:

- aged 75 years and over (57%) and aged 20 to 24 (69%);
- of Aboriginal (63%), Torres Strait Islander (65%) or both Aboriginal and Torres Strait Islander origin (60%);
- who usually lived in the Northern Territory (67%); and
- who usually lived in very remote areas (60%).

The lowest 2006-2011-2016 linkage rate by sub-population was those aged 75 years and over (29%) while the North Territory (48%) had the lowest linkage rate by state.

Most sub-populations followed a trend in their linkage rates across the three Census periods, although certain sub-populations fell considerably. Persons aged 15 to 19 in 2006 initially linked 74% of records to the 2011 Census, however dropped to 51% when linking to the 2016 Census. This is likely due to the high level of mobility as persons enter the 20 to 29 age range.

Traditionally, the Census Post Enumeration Survey (PES) has shown that the Census has higher rates of undercount for people of Aboriginal and/or Torres Strait Islander origin, those aged between 20 and 29 and for those in the Northern Territory. As expected, the lower ACLD linkage rates broadly aligned with the same groups that experience higher levels of undercount in the 2016 Census. One additional group that had lower linkage rates were persons aged 75 and over at the time of the 2006 Census who, due to age, had an increased risk of death over the ensuing ten years. Further information on Census undercount can be found in Census of Population and Housing: Details of Overcount and Undercount, 2011 (cat. no. 2940.0) and Census of Population and Housing - Details of Undercount, 2016 (cat. no. 2940.0).

Further data cubes demonstrating the linkage rates for various sub-populations are available as an attachment to this Information paper.

3.1 LINKAGE ACCURACY

The following quality measures were calculated for the ACLD and indicate a good level of overall quality:

1. The linkage rate, being the proportion of the 2006 ACLD Panel records linked to a 2011 Census record and then again to a 2016 Census record, including both true matches and false links.
2. The estimated proportion of correctly linked records, otherwise referred to as 'linkage precision'.
3. The consistency of reporting of common information between record pairs.

3.1.1 Linkage Precision

Not all record pairs assigned as links in a data linkage process are a true match, that is, a record pair belonging to the same individual. While the methodology is designed to ensure that the vast majority of links are true, some are actually false, i.e. the records in the link belong to different people rather than the same person. The linkage strategy used for the ACLD was designed to ensure a high level of accuracy while also achieving a sufficiently high number of links to enable longitudinal research. Accordingly, the strategy was restrictive and conservative.

One of the key measures of linkage quality is the proportion of links in the dataset that are false. The number of false links is able to be estimated through the use of methods such as clerically reviewing a sample of links, or by using modelling techniques. Once an estimate of the number of false links is obtained, a 'precision' can be calculated. The precision is an estimate of the proportion of links that are matches (i.e. belonging to the same entity).

$$Precision = \frac{Total\ links - False\ link\ estimate}{Total\ links}$$

Once the precision of the dataset is estimated, the false link rate is easily calculated.

$$False\ link\ rate = 1\text{-}\ Precision$$

With clerical review unavailable for the re-link of the 2006 panel, the model designed by Chipperfield et al (2018) known as the Feasibility Calculator (FC) was used as the primary method of calculating precision and setting a cut-off for the 2006-2011 re-link for the 2006 Panel sample. The FC uses the theory developed by Fellegi & Sunter (1969) to conduct a record linkage simulation multiple times in order to estimate precision. The FC then compiles the results of these simulated linkages to calculate the lowest linkage weight at the desired level of precision in each probabilistic linkage pass. These results can then be used to inform a single cut-off point for probabilistic linkage results. Due to the unavailability of name information the ability to distinguish a unique link becomes more difficult, so to ensure a high quality linkage while maintaining a high linkage rate it was decided to set the desired estimated cumulative precision at 95%, or an estimated false link rate of approximately 5%. This method achieved a 77.2% linkage rate when linking the 2006 Panel to 2011 Census records.

Precision estimation for the 2011-2016 linkage of the ACLD involved conducting clerical review on a stratified random sample of links. Potential links were stratified by their link weight value, with a minimum of 5% of links sampled from each individual link weight value (after rounding down to the nearest integer). The results of the clerical review were used to calculate precision estimates for links grouped by pass and rounded link weight value, which were then applied to the entire set of linkage results. This provided an estimate of precision for each individual link, which can be referred to as 'marginal precision', and is the likelihood of a single link being 'true' (i.e. the records belonging to the same person). Using the marginal precision, the 'cumulative precision' of the final set of one-to-one links could be estimated, i.e. the overall precision of the linked dataset.

After producing both marginal and cumulative precision estimates, a cut-off point was selected. This cut-off is intended to optimise both the number of links and cumulative precision of the links retained above the cut-off point, while at the same time maintaining a high level of marginal precision for every individual link above the cut-off. The marginal precision estimates were used to select the cut-off, with all links with a marginal precision of at least 81% being retained. This resulted in a final file of 605,626 links once the cut-off was applied, with an estimated cumulative precision of 98.6%, or a false link rate of 1.4%, for these links.

Clerical review relies upon judgment by a well-trained individual, therefore, while efforts are taken to minimise the risk, it is possible for a link to be incorrectly assigned as a match or non-match. The method for measuring precision developed by Chipperfield et al (2018) was used to provide an independent model-based estimate of the precision. While the clerical estimate of cumulative precision for the 2011-2016 linkage was 98.6%, the model-based approach estimated the precision to be over 99%. The precision as estimated by the clerical review process was retained as the more conservative estimate.

Table 2 provides a summary of the precision estimate and false link rate by the pass where each link was selected (estimated via clerical review) for the 2011-2016 linkage.

TABLE 2 - PRECISION ESTIMATES AND FALSE LINK RATES, By Pass Number, 2011-2016 linkage (2006 Panel)

Pass Number (a) (no.)	**Proportion of Overall Links** (%)	**Estimated True Link Rate / Precision Estimate** (%)	**Estimate False Link Rate** (%)
1	75.5	100	0
2	15.1	94.4	5.6
3	1.2	96.4	3.6
4	1.1	95.3	4.7
5	0.3	92.9	7.1
6	0.6	99.8	0.2
7	1.3	96.2	3.8
8	0.9	93.8	6.2
9	4.0	95.9	4.1
Total(b)	**100**	**98.6**	**1.4**

(a) Pass number 1 refers to the deterministic linkage.
(b) Data presented in the table have not been perturbed.

The conservative and restrictive nature of the blocking and linking strategy, accompanied by quality controls that were implemented during clerical review and the desired level of linkage precision, helped to minimise the estimated number of false links throughout the linkage process.

Over three quarters of all links were achieved in the first pass of the project (78.4% for 2006-2011 and 75.5% for 2011-2016), which used a deterministic linking methodology to identify and filter matches. This pass implemented tight geographic and demographic restrictions to maximise the number of high quality links assigned and to limit the amount of alternative comparisons required. Using this approach, links were only accepted if a single unique record pair was identified.

3.1.2 Consistency of Common Information on Record Pairs

In data linkage projects, geographic boundaries function as blocking variables that restrict the search for links to records which agree on the defined geography. They are also used as linking variables, and when combined with other linking fields (such as hashed name (2011-2016 only), age, sex and date of birth), they provide a high level of uniqueness, and reduce the likelihood of linking to an incorrect record.

Tables 3a and 3b display the number of records that had consistent information on key linking variables, grouped by levels of geography.

TABLE 3a - CONSISTENCY OF LINKED RECORDS, By Geography And Selected Linking

Fields, 2006 Panel, 2006-2011 linkage

	Consistency of key linkage fields(a)(b)(c) (no.)	(%)
MESH BLOCK		
Age exact, Mesh Block, Sex, DOB Day and Month agree	594,727	79.2
Age exact, Mesh Block, Sex agree	34,784	4.6
Age +/- 1 years, Mesh Block, Sex agree	20,190	2.7
STATISTICAL AREA LEVEL 2		
Age exact, SA2, Sex, DOB Day and Month agree	62,481	8.4
Age +/- 1 years, SA2, Sex, DOB Day and Month agree	412	0.1
Age +/- 1 years, SA2, Sex agree	16,513	2.2
STATISTICAL AREA LEVEL 4		
Age exact, SA4, Sex, DOB Day and Month agree	21,722	2.9
Total records included	751,199	99.2
Total records linked	**756,945**	**100**

(a) Only includes records that agree on all key linking fields.
(b) Categories are mutually exclusive. Records that agree in each category are excluded from subsequent categories.
(c) Percentages may not add up to the total due to rounding.

TABLE 3b - CONSISTENCY OF LINKED RECORDS, By Geography And Selected Linking Fields, 2006 Panel, 2011-2016 linkage

	Consistency of key linkage fields(a)(b)(c) (no.)	(%)
MESH BLOCK		
First name hash, Surname hash, Age exact, Mesh Block, Sex, DOB Day and Month agree	371,378	61.3
First name hash, Surname hash, Age exact, Mesh Block, Sex agree	93,577	15.5
Age exact, Mesh Block, Sex, DOB Day and Month agree	66,203	10.9
Age exact, Mesh Block, Sex agree	4,214	0.7
Age +/- 1 years, Mesh Block, Sex agree	15,853	2.6
STATISTICAL AREA LEVEL 2		
First name hash, Surname hash, Age +/- 1 years, SA2, Sex, DOB Day and Month agree	17,215	2.8
Age exact, Mesh Block, Sex, DOB Day and Month agree	4,589	0.8
Age +/- 1 years, SA2, Sex agree	3,234	0.5
STATISTICAL AREA LEVEL 4		
First name hash, Surname hash, Age +/- 1 years, SA4, Sex, DOB Day and Month agree	17,526	2.9
Age +/- 1 years, SA4, Sex, DOB Day and Month agree	4,149	0.7
Total records included	597,938	98.7
Total records linked	**605,618**	**100**

(a) Only includes records that agree on all key linking fields.
(b) Categories are mutually exclusive. Records that agree in each category are excluded from subsequent categories.
(c) Percentages may not add up to the total due to rounding.

Approximately 99% of all records that were matched in the ACLD linkage process agreed on small to medium levels of geographic area combined with other key linking fields, such as first name and surname hash codes (only for the 2011-2016 linkage), age, sex and date of birth. Analysis of consistency from the 2006 Census to the 2016 Census was not undertaken due to complexities in comparing geography. While the number of consistent fields can give a strong indication of likely linkage quality, other factors should be taken into account, for example, the expected number of people in a geographic area that are likely to share a characteristic by chance. A tolerance of plus or minus one year was used at certain parts of the linkage process to cater for persons who may have understated their age in one Census and/or overstated it in another Census or vice versa.

By contrast, record pairs may have inconsistent information and yet be a match. Inconsistent information may be recorded for the same person in different Censuses due to a range of factors, including:

- transcription errors in the Census, where the wrong category is selected or the information is transposed, such as the day the person was born being reported in the month field instead of in the day field;
- data capture errors, where the Census form is scanned using Optical Character Recognition (OCR) software and certain characters may be mis-classified, such as a 1 captured as a 7 or a 3 as an 8;
- reporting errors, where information is given for the wrong member of the household (e.g. person 1's information is reported for person 3) or where the person completing the Census form for a household guesses or estimates information about a fellow household member;
- information that was not stated by the respondent and has been imputed as part of Census processing (such as age or sex), while set to missing for linking, the imputed values are included in the analytical dataset;
- census form questions are interpreted differently at each Census; or
- questions are coded differently for each Census.

Of particular note is inconsistency due to non-reporting of name and date of birth in the 2011 Census and the 2016 Census. Respondents are becoming less likely to provide their date of birth, with 90% reporting in the 2011 Census decreasing to 81% reported date of birth in the 2016 Census. Further, just over one per cent of Australians had a missing, or blank, response for first name or surname in the 2016 Census. There appeared to be a relationship between having a missing response for both first name and surname and non-response on other variables. Of the people who did not report first name and surname, approximately half did not report at least one of sex, age, or Indigenous status. The vast majority of missing responses came from paper forms, with the overall level of missing responses in the 2016 Census remaining low.

3.1.3 Comparison with the original 2006 Panel linkage

Table 4 compares the final results of the original 2006 Panel linkage with the revised linkage.

TABLE 4 – COMPARISON OF LINKAGE RESULTS, 2006-2011

	Original linkage (2006-11)	Re-link (2006-11)
Linked records (no.)	800,758	756,945
Linkage rate (%)	82.0	77.2
Precision	Approx. 90-95%	95%

While the linkage rate has reduced there is greater confidence in the precision of the links that have been achieved in the re-link of the 2006 Panel due to the enhanced linking methodology implemented for the linkage. Over the entire panel 81.6% of records always achieved the same result (same link identified, or not linked). The changes in links can be viewed in Table 5.

TABLE 5 – STATUS OF LINKS, 2006-2011

	2006 Panel records (no.)	2006 Panel records (%)
Same link	670,073	68.4
Different link	37,206	3.8
New link	49,666	5.1
Lost link	93,479	9.5
Never linked	129,238	13.2

(a) Data presented in the table have not been perturbed.

3.2 CHARACTERISTICS OF LINKED AND UNLINKED 2006 ACLD PANEL SAMPLE

The random sample selected from the 2006 Census for the 2006 ACLD Panel was designed to be representative of the Australian population by age, sex and jurisdiction as well as other characteristics such as Indigenous status and country of birth.

Table 6 shows the distribution of key populations across the 2006 Census, the 2006 ACLD Panel sample and the 2006-2011 linked results.

TABLE 6 - SELECTED CHARACTERISTICS, By 2006 Census, 2006 ACLD Panel Sample, 2006-2011 ACLD Linked Results

	2006 Census		2006 Panel Sample		Linked Results (2006-2011)		Weighted Linked Results (a) (2006-2011)	
	(no.)	(%)	(no.)	(%)	(no.)	(%)	(no.)	(%)
SEX								
Male	9 896 500	49.3	480 285	49.0	364 727	48.2	9 085 806	49.3
Female	10 165 146	50.7	499 372	51.0	392 222	51.8	9 326 779	50.7
STATE/TERRITORY OF USUAL RESIDENCE								
New South Wales	6 549 174	32.6	323 136	33.0	250 070	25.5	6 037 632	32.8
Victoria	4 932 422	24.6	244 095	24.9	191 981	19.6	4 620 771	25.1
Queensland	3 904 531	19.5	192 606	19.7	144 427	14.7	3 611 492	19.6
South Australia	1 514 340	7.5	75 481	7.7	59 386	6.1	1 384 197	7.5
Western Australia	1 959 088	9.8	95 795	9.8	73 948	7.5	1 843 423	10.0
Tasmania	476 481	2.4	23 787	2.4	18 624	1.9	428 564	2.3
Northern Territory	192 899	1.0	8 469	0.9	5 573	0.6	189 052	1.0
Australian Capital Territory	324 034	1.6	16 186	1.7	12 866	1.3	295 676	1.6
AGE GROUP								
0-9	2 579 496	12.9	127 331	13.0	92 684	12.2	2 586 620	14.0

10-19	2 756 102	13.7	132 937	13.6	97 586	12.9	2 547 772	13.8
20-29	2 684 371	13.4	128 760	13.1	91 875	12.1	2 650 903	14.4
30-39	2 893 058	14.4	140 271	14.3	113 887	15.0	2 891 972	15.7
40-49	2 942 353	14.7	142 911	14.6	120 932	16.0	2 875 902	15.6
50-59	2 574 589	12.8	126 285	12.9	107 379	14.2	2 412 469	13.1
60-69	1 733 297	8.6	86 385	8.8	72 041	9.5	1 514 365	8.2
70-79	1 168 675	5.8	58 277	5.9	43 486	5.7	755 872	4.1
80 and over	729 705	3.6	36 502	3.7	17 071	2.3	176 533	1.0
INDIGENOUS STATUS								
Non-Indigenous	18 266 814	91.1	942 253	96.2	733 032	96.8	17 654 813	95.9
Aboriginal and/or Torres Strait Islander	455 027	2.3	21 985	2.2	13 892	1.8	524 075	2.8
Aboriginal	407 700	2.0	19 697	2.0	12 449	1.6	466 722	2.5
Torres Strait Islander	29 515	0.1	1 449	0.1	940	0.1	38 103	0.2
Both Aboriginal and Torres Strait Islander	17 812	0.1	839	0.1	503	0.1	19 250	0.1
Not stated	1 133 449	5.6	15 416	1.6	10 020	1.3	233 603	1.3
Total (b)(c)(d)	**20 061 646**	**100**	**979 662**	**100**	**756 945**	**100**	**18 412 584**	**100**

(a) For more information on weighting see chapter 3.4.
(b) Data presented in the table have been perturbed. As a result the sum of individual categories may not align with totals.
(c) Includes Other Territories.
(d) Includes Migratory areas.

The distribution of the ACLD file by sub-population was generally well aligned with both the 2006 Panel sample and the entire 2006 Census. When looking at the relative difference between these proportions, however, some differences are more clearly observed.

Compared with the entire 2006 Census, the linked 2006 ACLD Panel contains relatively more records for people aged 40-49 and 50-59 years, and to a lesser extent those aged 60-69 years. By contrast, the linked 2006 Panel contains relatively fewer records for people aged 20-29 years and 80 years and over. This is applicable for both the 2006-2011 and 2006-2011-2016 linkages, with the latter having increased proportional differences when compared to the 2006 Census.

In general, the distribution of weighted counts for the linked ACLD file is close to that of the entire 2006 Census, but it should be noted that the weighting process is not designed to produce counts corresponding to the population in 2006. Rather, the weighted population is that of people who were in scope of both the 2006 and 2011 Censuses for the 2006-2011 linkage and of people who were in scope of the 2006, 2011 and 2016 Censuses for the 2006-2011-2016 linkage (see Section 3.4 Weighting). Thus, for example, the lower proportion of older people in the linked file, even after weighting, reflects the impact on the 2006 Panel sample of deaths that occurred between 2006, 2011 and 2016.

Further data cubes demonstrating more detailed population distributions are provided as an attachment to this Information paper.

3.3 REASONS FOR UNLINKED RECORDS

There are two main reasons why records from the 2006 Panel sample were not linked to a 2011 Census and/or 2016 Census record:

- records belonging to the same individual were present in the 2006 Panel sample and the 2011 and/or 2016 Census but these records failed to be linked because they contained missing or inconsistent information; or
- there was no 2011 or 2016 Census record corresponding to the 2006 Panel sample record because the person was not counted in the 2011 and/or 2016 Census.

3.3.1 Missing and/or inconsistent information

In these cases, the true match was present in the pool of all record pairs but it was not identified because there was a high level of inconsistency between information on each Census, or key linking fields were missing altogether. The reasons for the match being missed can be categorised into the following groups:

- the missing or inconsistent information did not allow the record pair to be compared in the same blocking categories and could not be linked;
- the record pair did not contain enough unique common information to distinguish the match from other potential record pairs;
- the record pair was linked, but was attributed a low link weight as it contained a lot of missing or inconsistent information and was positioned below the cut-off identified in sample clerical review or modelling via the Feasibility Calculator; or
- the record pair was subjected to clerical review, but the high level of inconsistency did not enable it to be deemed a true link.

Accurate address coding was crucial in narrowing the search and differentiating between true and false links. It was a particular challenge for persons who had moved, since linkage was then heavily dependent on accurate recall and detailed information supplied in the 2011 and 2016 Censuses about the person’s address five years previous. Processing for the 2011 and 2016 Census involved coding for address five years ago to a fine level of geography, ideally Mesh Block. This was not always possible, due to insufficient and/or incorrect address information being supplied for some persons, potentially due to recall issues.

3.3.2 No 2011 or 2016 Census Record

A person included in the 2006 Panel sample may have had no equivalent 2011 Census and/or 2016 Census record because they were no longer in scope for the Census due to migration from Australia, or death between 2006 and 2016, or they may simply have been missed in the Census. If a 2006 Panel record was not linked to the 2011 Census then that person did not have the opportunity to be linked to the 2016 Census, due to these only being linked via the linked 2011 Census records.

According to mortality data compiled by the ABS from data supplied by the Registrars of Births, Deaths and Marriages, approximately 700,000 people died in Australia between 2006 and 2011 and approximately 913,000 between 2011 and 2016. If 5% of these people were selected in the 2006 Panel sample, then it could be estimated that up to 35,000 people could not have been linked due to death between 2006 and 2011. Similarly, migration data estimates that just over one million people left Australia as permanent emigrants between 2006 and 2011, while just over 1.4 million people left between 2011 and 2016, potentially resulting in up to 50,000 people from the 2006 Panel sample being unlikely to have a corresponding 2011 Census record. For more information please refer to the relevant releases of Migration, Australia (cat. no. 3412.0) and Deaths, Australia (cat. no. 3302.0).

Due to the size and complexity of the Census, it is inevitable that some people are missed and some are counted more than once. It is for this reason that the Census Post Enumeration Survey (PES) is run shortly after each Census, to provide an independent measure of Census coverage. The PES determines how many people should have been counted in the Census, how many were missed (undercount), and how many were counted more than once (overcount). It also provides information on the characteristics of those in the population who have been under- or overcounted.

The net undercount rate was 1.7% for the 2011 Census and 1% for the 2016 Census, with higher rates for Aboriginal and Torres Strait Islander people than for the non-Indigenous population. Thus approximately 15,000 people from the 2006 Panel sample could have been missed in the 2011 Census. This estimate is a starting point only and does not take into account the likelihood of people being missed in successive Censuses. For more information please refer to Census of Population and Housing - Details of Undercount, 2011 (cat. no. 2940.0) for the 2011 Census and Census of Population and Housing: Details of Overcount and Undercount, 2016 (cat. no. 2940.0) for the 2016 Census.

When taking into account all of these factors, it is estimated that nearly half of the unlinked 2006 Panel sample (100,000 out of the 222,717 unlinked records) would not have a corresponding record in the 2011 Census. This would indicate that the initial linkage rate of 77% for the 2006-2011 linkage could be representative of up to 89% of the population that actually had an opportunity to be linked.

The proportion of links achieved in the 2011-2016 linkage of the 2006 Panel is approximately 3.5% of the 2011 Census population. Using the factors outlined above, approximately 913,000 people died between 2011 and 2016, therefore it could be estimated that almost 32,000 people could not have been linked due to death between 2011 and 2016. Similarly, migration data estimates that just over 1.4 million people left Australia as permanent emigrants between 2011 and 2016, potentially resulting in approximately 49,000 people being unlikely to have a corresponding 2016 Census record due to migration between 2011 and 2016. For more information please refer to the relevant releases of Migration, Australia (cat. no. 3412.0) and Deaths, Australia (cat. no. 3302.0). Taking into account the net undercount rate of 1% for the 2016 Census, it is estimated that almost 8,000 persons may have been missed and therefore missing a corresponding 2016 Census record.

Therefore it is estimated that almost 59% of the unlinked 2011 Census records from the 2006 Panel sample (89,000 of 151,000 unlinked records) would not have had a corresponding records in the 2016 Census. This would indicate that the initial linkage rate of 80% could be representative of almost 92% of the population that actually had an opportunity to be linked.

Thus it is estimated that 50% of the unlinked records from the 2006 Panel (189,000 of 374,000 unlinked records) would not have a corresponding record in the 2016 Census, however this estimate does not take into account persons that were out of scope or missed for the 2011 Census but may have come back into scope for the 2016 Census. This would indicate that the 62% linkage rate for the 2006 Panel that linked to both the 2011 Census and the 2016 Census could be representative of approximately 81% of the population that actually had an opportunity to be linked from 2006 through to 2016.

3.4 WEIGHTING

Weighting is the process of adjusting a sample to infer results for the relevant population. To do this, a 'weight' is allocated to each sample unit - in this case, persons. The weight can be considered an indication of how many people in the relevant population are represented by each person in the sample. In the case of the ACLD, populations are defined in terms of a set of Censuses. The 2006-2011 longitudinal population is defined as those people in scope of the 2006 and 2011 Censuses while the 2006-2011-2016 longitudinal population is those people in scope of the 2006, 2011 and 2016 Censuses. The longitudinal weights were created for linked records in the ACLD to enable longitudinal population estimates to be produced. Cross-sectional population estimates for 2006, 2011 and 2016 are available from each Census.

The 2006 Panel of the ACLD is a random sample of 5% of the 2006 Census. As such, each person in the sample should represent about 20 people in the 2006 Census population. Between Censuses, however, the in scope population changes as people die or move overseas. In addition, Census net undercount and data quality can affect the capacity to link equivalent records across waves. The weights of the linked records on the ACLD were calibrated to the estimated population that was in scope of the 2006 and 2011 Censuses and then again for the 2006, 2011 and 2016 Censuses. The weights were based on four components: the design weight, undercoverage adjustment, missed link adjustment and population benchmarking.

Two distinct weights were designed to allow for analysis of either the 2006-2011 or 2006-2011-2016 longitudinal population. Unique weights have not been designed for the 2011-2016 longitudinal population that have been linked within the 2006 Panel. It is advised that analysis of this particular population should be undertaken using the 2011 Panel, which was designed to be representative of the 2011 Census population.

The mean final weights for the linked records is 25.0 for females and 26.6 for males in the 2006-2011 longitudinal population and 29.4 for females and 31.5 for males in the 2006-2011-2016 longitudinal population. The weights range between 16.05 and 176.9 for 2006-2011 and between 15.9 and 341.3 for 2006-2011-2016. The mean weight was higher for Aboriginal and Torres Strait Islander persons and for people in the Northern Territory.

The 2006-2011 and 2006-2011-2016 longitudinal population benchmarks are based on the 2011 and 2016 Estimated Resident Population (ERP), which is adjusted by the estimated probability a person belongs to the longitudinal population. This probability is formed using the Census reported address five year ago variable from the 2011 or 2016 Census. Further information on this approach can be found in the paper Chipperfield, Brown & Watson (2016). See References section for details of this publication.

For more information about weighting please refer to the Appendix.

ACLD 2011-16

PRODUCT OVERVIEW 2011-16

The 2011-16 ACLD is a representative sample of over 1.2 million records from the 2011 Census (Wave 2) brought together with corresponding records from the 2016 Census (Wave 3). The 2011 Panel includes new births and migrants since the 2006 Census, and is a rich source for exploring how Australian society has changed between the 2011 and 2016 Censuses.

The 2011-16 ACLD product is recommended for analysis of the 2011-16 longitudinal population.

Linkage Results

3. LINKAGE RESULTS, 2011-2016, 2011 PANEL

At the completion of the linkage process 927,520 (76%) of the 1,221,057 records from the 2011 ACLD Panel sample were linked to a 2016 Census record to create the linked 2011-2016 ACLD file with an estimated false link rate of 1.4%.

All results presented in this publication (unless identified in the relevant table) are based on characteristics from the 2011 ACLD Panel sample and have been confidentialised to prevent the identification of individuals.

Table 1 displays the linkage rate for a range of sub-populations.

TABLE 1 - LINKAGE RATES, By Selected Characteristics

	2011 Panel sample (no.)	**Linked records** (no.)	**Linkage rate** (%)
SEX			
Male	600 724	450 092	74.9
Female	620 334	477 426	77.0
AGE GROUP			
0-14	236 383	189 641	80.2
15-19	79 971	57 114	71.4
20-24	82 222	52 044	63.3
25-29	85 198	57 331	67.3
30-39	168 979	127 974	75.7
40-49	172 576	139 142	80.6
50-59	155 652	127 702	82.0
60-69	121 036	99 537	82.2
70-74	40 657	32 211	79.2
75 and over	78 384	44 823	57.2
INDIGENOUS STATUS			
Non-Indigenous	1 171 794	897 076	76.6
Aboriginal	29 156	18 515	63.6
Torres Strait Islander	1 819	1 174	64.5
Both Aboriginal and Torres Strait Islander	1 243	802	64.6
Not stated	17 050	9 948	58.3
STATE/TERRITORY OF USUAL RESIDENCE			
New South Wales	393 519	298 795	75.9
Victoria	304 513	233 623	76.7
Queensland	245 366	183 703	74.9
South Australia	91 555	71 650	78.3
Western Australia	125 449	95 053	75.8
Tasmania	28 580	21 831	76.4
Northern Territory	11 628	7 240	62.3
Australian Capital Territory	20 272	15 530	76.6
REMOTE AREAS			
Major Cities	852 825	651 866	76.4
Inner Regional	228 174	174 567	76.5
Outer Regional	110 441	82 485	74.7
Remote	16 570	11 462	69.2
Very Remote	10 201	6 016	59.0
No Usual Address	2 593	1 002	38.6
Total(a)(b)(c)	**1 221 057**	**927 520**	**76.0**

(a) Data presented in the table have been perturbed. As a result, the sum of individual categories may not align with totals.
(b) Includes Other Territories.
(c) Includes Migratory areas.

The linkage rates for the 2011-2016 ACLD were relatively consistent across most sub-populations and were in line with expected results. Compared with the overall linkage rate of 76%, the sub-populations which achieved the highest linkage rates were persons:

- aged 60 to 69 years (82%), followed by 50 to 59 years (82%) and 0 to 14 years (80%);

- of non-Indigenous origin (77%);
- who usually lived in South Australia (78%); and
- who usually lived in major cities (76%) and inner regional areas (77%).

The sub-populations which achieved the lowest linkage rates were persons:

- aged 20-24 years (63%) and 75 years and over (57%);
- of Aboriginal (64%), Torres Strait Islander (65%) or both Aboriginal and Torres Strait Islander origin (65%);
- who usually lived in the Northern Territory (62%); and
- who usually lived in remote (69%) and very remote areas (59%) or who had no usual address in 2011 (39%).

Traditionally, the Census Post Enumeration Survey (PES) has shown that the Census has higher rates of undercount for people of Aboriginal and/or Torres Strait Islander origin, those aged between 20 and 29 and for those in the Northern Territory. As expected, the lower ACLD linkage rates broadly aligned with the same groups that experience higher levels of undercount in the 2016 Census. One additional group that had lower linkage rates were persons aged 75 and over at the time of the 2011 Census who, due to age, had an increased risk of death over the ensuing five years. Further information on Census undercount can be found in Census of Population and Housing: Details of Overcount and Undercount, 2016 (cat. no. 2940.0).

Further, data cubes demonstrating the linkage rates for various sub-populations are available as an attachment to this Information paper.

3.1 LINKAGE ACCURACY

The following quality measures were calculated for the ACLD and indicate a good level of overall quality:

1. The linkage rate, being the proportion of the 2011 ACLD Panel records linked to a 2016 Census record, including both true matches and false links.
2. The estimated proportion of correctly linked records, otherwise referred to as 'linkage precision'.
3. The consistency of reporting of common information between record pairs.

3.1.1 Linkage Precision

Not all record pairs assigned as links in a data linkage process are a true match, that is, a record pair belonging to the same individual. While the methodology is designed to ensure that the vast majority of links are true, some are actually false, i.e. the records in the link belong to different people rather than the same person. The linkage strategy used for the ACLD was designed to ensure a high level of accuracy while also achieving a sufficiently high number of links to enable longitudinal research. Accordingly, the strategy was restrictive and conservative.

One of the key measures of linkage quality is the proportion of links in the dataset that are false. The number of false links is able to be estimated through the use of methods such as clerically reviewing a sample of links, or by using modelling techniques. Once an estimate of the number of false links is obtained, a 'precision' can be calculated. The precision is an estimate of the proportion of links that are matches (i.e. belonging to the same entity).

$$Precision = \frac{Total\ links - False\ link\ estimate}{Total\ links}$$

Once the precision of the dataset is estimated, the false link rate is easily calculated.

$$\textit{False link rate} = 1 - \textit{Precision}$$

Precision estimation for the ACLD involved conducting clerical review on a stratified random sample of links. Potential links were stratified by their link weight value, with a minimum of 5% of links sampled from each individual link weight value (after rounding down to the nearest integer). After reviewing the sample, the results were used to calculate precision estimates for links grouped by pass and rounded link weight value. These estimates were then applied to the entire set of linkage results. This provided an estimate of precision for each individual link, which can be referred to as 'marginal precision'. Using the marginal precision, the 'cumulative precision' of the final set of one-to-one links could be estimated.

After producing both marginal and cumulative precision estimates, a cut-off point was selected. This cut-off is intended to optimise both the number of links and cumulative precision of the links retained above the cut-off point, while at the same time maintaining a high level of marginal precision for every individual link above the cut-off. The marginal precision estimates were used to select the cut-off, with all links with a marginal precision of at least 81% being retained. This resulted in a final file of 927,520 links once the cut-off was applied, with an estimated cumulative precision of 98.6%, or a false link rate of 1.4%, for these links.

Clerical review relies upon judgment by a well trained individual, therefore, while efforts are taken to minimise the risk, it is possible for a link to be incorrectly assigned as a match or non-match. An alternative way of measuring precision is through the use of models. We applied the method of Chipperfield et al (2018) to provide an independent model-based estimate of the precision. While the clerical estimate of cumulative precision was 98.6%, the model-based approach estimated the precision to be over 99%. The precision as estimated by the clerical review process was retained as the more conservative estimate.

Table 2 provides a summary of the precision estimate and false link rate by the pass where each link was selected (estimated via clerical review).

TABLE 2 - PRECISION ESTIMATES AND FALSE LINK RATES, By Pass Number

Pass Number (a)	Proportion of Overall Links	Estimated True Link Rate / Precision Estimate	Estimate False Link Rate
(no.)	(%)	(%)	(%)
1	72.7	100	0
2	15.7	94.4	5.6
3	1.2	96.4	3.6
4	1.5	95.3	4.7
5	0.8	92.9	7.1
6	1.1	99.8	0.2
7	1.6	96.2	3.8
8	1.0	93.8	6.2
9	4.4	95.9	4.1
Total(b)	**100**	**98.6**	**1.4**

(a) Pass number 1 refers to the deterministic linkage.
(b) Data presented in the table have been unperturbed.

The conservative and restrictive nature of the blocking and linking strategy, accompanied by quality controls that were implemented during clerical review, helped to minimise the estimated number of false links throughout the linkage process.

Almost three quarters (73%) of all links were achieved in the first pass of the project, which used a deterministic linking methodology to identify and filter matches. This pass implemented tight geographic and demographic restrictions to maximise the number of high quality links assigned and to limit the amount of alternative comparisons required. Using this approach, links were only accepted if a single unique record pair was identified.

3.1.2 Consistency of Common Information on Record Pairs

In data linkage projects, geographic boundaries function as blocking variables that restrict the search for links to records which agree on the defined geography. They are also used as linking variables, and when combined with other linking fields (such as hashed name, age, sex and date of birth), they provide a high level of uniqueness, and reduce the likelihood of linking to an incorrect record.

Table 3 displays the number of records that had consistent information on key linking variables, grouped by levels of geography.

TABLE 3 - CONSISTENCY OF LINKED RECORDS, By Geography And Selected Linking Fields

	Consistency of key linkage fields(a)(b)(c)	
	(no.)	(%)
MESH BLOCK		
First name hash, Surname hash, Age exact, Mesh Block, Sex, DOB Day and Month agree	530,305	57.2
First name hash, Surname hash, Age exact, Mesh Block, Sex agree	160,953	18.3
Age exact, Mesh Block, Sex, DOB Day and Month agree	96,202	10.4
Age exact, Mesh Block, Sex agree	7,176	0.8
Age +/- 2 years, Mesh Block, Sex agree	31,223	3.4
STATISTICAL AREA LEVEL 2		
First name hash, Surname hash, Age +/- 2 years, SA2, Sex, DOB Day and Month agree	28,767	3.1
Age exact, Mesh Block, Sex, DOB Day and Month agree	8,677	0.9
Age +/- 2 years, SA2, Sex agree	7,226	0.8
STATISTICAL AREA LEVEL 4		
First name hash, Surname hash, Age +/- 2 years, SA4, Sex, DOB Day and Month agree	33,103	3.6
Age +/- 2 years, SA4, Sex, DOB Day and Month agree	8,103	0.9
Total records included	911,735	98.3
Total records linked	**927,520**	**100**

(a) Only includes records that agree on all key linking fields.
(b) Categories are mutually exclusive. Records that agree in each category are excluded from subsequent categories.
(c) Percentages may not add up to the total due to rounding.

Over 98% of all records that were matched in the ACLD linkage process agreed on small to medium levels of geographic area combined with other key linking fields, such as first name and surname hash codes, age, sex and date of birth. While the number of consistent fields can give a strong indication of likely linkage quality, other factors should be taken into account, for example, the expected number of people in a geographic area that are likely to share a characteristic by chance. A tolerance of plus or minus one year was used at certain parts of the linkage process to cater for persons who may have understated their age in 2011 and/or overstated it in 2016 or vice versa.

By contrast, record pairs may have inconsistent information and yet be a match. Inconsistent information may be recorded for the same person in different Censuses due to a range of factors, including:

- transcription errors in the Census, where the wrong category is selected or the information is transposed, such as the day the person was born being reported in the month field instead of in the day field;
- data capture errors, where the Census form is scanned using Optical Character Recognition (OCR) software and certain characters may be mis-classified, such as a 1 captured as a 7 or a 3 as an 8;
- reporting errors, where information is given for the wrong member of the household (e.g. person 1's information is reported for person 3) or where the person completing the Census form for a household guesses or estimates information about a fellow household member;
- information that was not stated by the respondent and has been imputed as part of Census processing (such as age or sex), while set to missing for linking, the imputed values are included in the analytical dataset;
- census form questions are interpreted differently at each Census; or
- questions are coded differently for each Census.

Of particular note is inconsistency due to non-reporting of name and date of birth. Respondents are becoming less likely to provide their date of birth, with 90% reporting in the 2011 Census decreasing to 81% reported date of birth in the 2016 Census. Further, just over one per cent of Australians had a missing, or blank, response for first name or surname in the 2016 Census. There appeared to be a relationship between having a missing response for both first name and surname and non-response on other variables. Of the people who did not report first name and surname, approximately half did not report at least one of sex, age, or Indigenous status. The vast majority of missing responses came from paper forms, with the overall level of missing responses in the 2016 Census remaining low.

3.2 CHARACTERISTICS OF LINKED AND UNLINKED 2011 ACLD PANEL SAMPLE

The random sample selected from the 2011 Census for the 2011 ACLD Panel was designed to maximise overlap with the 2006 ACLD Panel, while also being representative of the Australian population by age, sex and jurisdiction as well as other characteristics such as Indigenous status and country of birth. The 2011 Panel sample size was increased in comparison to the 2006 Panel sample size primarily due to the increase in the Australian population from 2006 to 2011. The 2011 Panel size was increased slightly to 5.7%, to achieve a linked sample size closer to 5% of the population after allowing for missed links and people no longer being in scope of the ACLD due to death or overseas migration**.**

Table 4 shows the distribution of key populations across the 2011 Census, the 2011 ACLD Panel sample and the linked results.

TABLE 4 - SELECTED CHARACTERISTICS, By 2011 Census, 2011 ACLD Panel Sample, ACLD Linked Results

	2011 Census		2011 Panel Sample		Linked Results		Weighted Linked Results (a)	
	(no.)	(%)	(no.)	(%)	(no.)	(%)	(no.)	(%)
SEX								
Male	10 634 012	49.4	600 724	49.2	450 092	48.5	10 440 753	49.5
Female	10 873 706	50.6	620 334	50.8	477 426	51.5	10 639 417	50.5

STATE/TERRITORY OF USUAL RESIDENCE								
New South Wales	6 917 656	32.2	393 519	32.2	298 795	32.2	6 787 716	32.2
Victoria	5 354 039	24.9	304 513	24.9	233 623	25.2	5 304 805	25.2
Queensland	4 332 727	20.2	245 366	20.1	183 703	19.8	4 223 043	20.0
South Australia	1 596 569	7.4	91 555	7.5	71 650	7.7	1 548 407	7.3
Western Australia	2 239 171	10.4	125 449	10.3	95 053	10.2	2 182 402	10.4
Tasmania	495 351	2.3	25 580	2.3	21 831	2.4	476 403	2.3
Northern Territory	211 943	1.0	11 628	1.0	7 240	0.8	211 411	1.0
Australian Capital Territory	357 218	1.7	20 272	1.7	15 530	1.7	343 595	1.6
AGE GROUP								
0-9	2 772 971	12.9	157 597	12.9	126 844	13.7	2 823 442	13.4
10-19	2 776 848	12.9	158 761	13.0	119 912	129	2 822 767	13.4
20-29	2 973 916	13.8	167 423	13.7	109 375	11.8	3 047 805	14.5
30-39	2 973 913	13.8	168 979	13.8	127 974	13.8	2 987 460	14.2
40-49	3 047 023	14.2	172 576	14.1	139 142	15.0	3 050 851	14.5
50-59	2 744 653	12.8	155 652	12.7	127 702	13.8	2 718 221	12.9
60-69	2 125 435	9.9	121 036	9.9	99 537	10.7	2 051 448	9.7
70-79	1 253 349	5.8	71 658	5.9	54 430	5.9	1 098 356	5.2
80 and over	839 609	3.9	47 387	3.9	22 603	2.4	479 854	2.3
INDIGENOUS STATUS								
Non-Indigenous	19 900 765	92.5	1 171 794	96.0	897 076	96.7	20 228 715	96.0
Aboriginal and/or Torres Strait Islander	548 368	2.5	32 218	2.6	20 491	2.2	617 382	2.9
Aboriginal	495 754	2.3	29 156	2.4	18 515	2.0	558 748	2.7
Torres Strait Islander	31 407	0.1	1 819	0.1	1 174	0.1	34 407	0.2
Both Aboriginal and Torres Strait Islander	21 205	0.1	1 243	0.1	802	0.1	24 227	0.1
Not stated	1 058 585	4.9	17 050	1.1	9 948	1.1	233 961	1.1
Total (b)(c)(d)	**21 507 719**	**100**	**1 221 057**	**100**	**927 520**	**100**	**21 080 214**	**100**

(a) For more information on weighting see chapter 3.4.
(b) Data presented in the table have been perturbed. As a result the sum of individual categories may not align with totals.
(c) Includes Other Territories.
(d) Includes Migratory areas.

The distribution of the ACLD file by sub-population was generally well aligned with both the 2011 Panel sample and the entire 2011 Census. When looking at the relative difference between these proportions, however, some differences are more clearly observed.

Compared with the entire 2011 Census, the linked 2011 ACLD Panel contains relatively more records for people aged 50-59 years, and to a lesser extent those aged 0-9 years, 40-49 years and 60-69 years. By contrast, the linked 2011 Panel contains relatively fewer records for people aged 20-29 years and 80 years and over. This is consistent with the 2006-2011 ACLD linkage as these subpopulations followed similar linkage rates.

In general, the distribution of weighted counts for the linked ACLD file is close to that of the entire 2011 Census, but it should be noted that the weighting process is not designed to produce counts corresponding to the population in 2011. Rather, the weighted population is that of people who were in scope of both the 2011 and 2016 Censuses (see Section 3.4 Weighting). Thus, for example, the lower proportion of older people in the linked file, even after weighting, reflects the impact on the 2011 Panel sample of deaths that occurred between 2011 and 2016.

Further data cubes demonstrating more detailed population distributions are provided as an attachment to this Information paper.

3.3 REASONS FOR UNLINKED RECORDS

There are two main reasons why records from the 2011 Panel sample were not linked to a 2016 Census record:

1. records belonging to the same individual were present in the 2011 Panel sample and the 2016 Census but these records failed to be linked because they contained missing or inconsistent information; or
2. there was no 2016 Census record corresponding to the 2011 Panel sample record because the person was not counted in the 2016 Census.

3.3.1 Missing and/or inconsistent information

In these cases, the true match was present in the pool of all record pairs but it was not identified because there was a high level of inconsistency between information on the 2011 ACLD Panel sample record and the 2016 Census record, or key linking fields were missing altogether. The reasons for the match being missed can be categorised into the following groups:

- the missing or inconsistent information did not allow the record pair to be compared in the same blocking categories and could not be linked;
- the record pair did not contain enough unique common information to distinguish the match from other potential record pairs;
- the record pair was linked, but was attributed a low link weight as it contained a lot of missing or inconsistent information and was positioned below the cut-off identified in sample clerical review; or
- the record pair was subjected to clerical review, but the high level of inconsistency did not enable it to be deemed a true link.

Accurate address coding was crucial in narrowing the search and differentiating between true and false links. It was a particular challenge for persons who had moved, since linkage was then dependent on the information supplied in 2016 about the person's address in 2011. Processing for the 2016 Census involved coding for address five years ago to a fine level of geography, ideally Mesh Block. This was not always possible, due to insufficient and/or incorrect address information being supplied for some persons, potentially due to recall issues.

3.3.2 No 2016 Census record

A person included in the 2011 ACLD Panel sample may have had no equivalent 2016 Census record because they were no longer in scope for the Census due to migration from Australia, or death between 2011 and 2016, or they may have been missed in the 2016 Census.

According to mortality data compiled by the ABS from data supplied by the Registrars of Births, Deaths and Marriages, approximately 913,000 people died in Australia between 2011 and 2016. If 5% of these people were selected in the 2011 Panel sample, then it could be estimated that up to 46,000 people could not have been linked due to death between 2011 and 2016. Similarly, migration data estimates that just over 1.4 million people left Australia as permanent emigrants over the same period, potentially resulting in up to 70,000 people from the 2011 Panel sample being unlikely to have a corresponding 2016 Census record. For more information please refer to the relevant releases of Migration, Australia (cat. no. 3412.0) and Deaths, Australia (cat. no. 3302.0).

Due to the size and complexity of the Census, it is inevitable that some people are missed and some are counted more than once. It is for this reason that the Census Post Enumeration Survey (PES) is run shortly after each Census, to provide an independent measure of Census coverage. The PES determines how many people should have been counted in the Census, how many were missed (undercount), and how many were counted more than once (overcount). It also provides information on the characteristics of those in the population who have been under- or overcounted.

The net undercount rate for the 2016 Census was 1%, with a higher rate for Aboriginal and Torres Strait Islander people than for the non-Indigenous population. Thus approximately 12,000 people from the 2011 Panel sample could have been missed in the 2016 Census. This estimate is a starting point only and does not take into account the likelihood of people being missed in successive Censuses. For more information please refer to Census of Population and Housing: Details of Overcount and Undercount, 2016 (cat. no. 2940.0).

When taking into account all of these factors, it is estimated that approximately 40% of the unlinked 2011 ACLD Panel sample (128,000 out of the 293,000 unlinked records) would not have a corresponding record in the 2016 Census. This would indicate that the initial linkage rate of 76% could be representative of up to 85% of the population that actually had an opportunity to be linked.

3.4 WEIGHTING

Weighting is the process of adjusting a sample to infer results for the relevant population. To do this, a 'weight' is allocated to each sample unit - in this case, persons. The weight can be considered an indication of how many people in the relevant population are represented by each person in the sample. Weights were created for linked records in the ACLD to enable longitudinal population estimates to be produced. Cross-sectional population estimates for 2011 and 2016 are available from each Census.

The 2011 Panel of the ACLD is a random sample of 5% of the Australian population in 2011. As such, each person in the sample should represent about 20 people in the population. Between Censuses, however, the in scope population changes as people die or move overseas. In addition, Census net undercount and data quality can affect the capacity to link equivalent records across waves. The weights of the linked records on the ACLD were calibrated to the estimated population that was in scope of both the 2011 and 2016 Censuses, 21,080,214 persons. The weights were based on four components: the design weight, undercoverage adjustment, missed link adjustment and population benchmarking.

The mean final weight for the linked records is 22.3 for females and 23.2 for males. The weights range between 14.8 and 83. The mean weight was higher for Aboriginal and Torres Strait Islander persons and for people in the Northern Territory.

The population benchmark is based on the 2016 Estimated Resident Population (ERP), which is adjusted by the estimated probability a person was also in Australia in 2011. This probability is formed using the 2016 Census reported address five year ago variable. Further information on this approach can be found in the paper Chipperfield, Brown & Watson (2016). See References section for details of this publication.

For more information about weighting please refer to the Appendix.

ACLD 2006-11

PRODUCT OVERVIEW 2006- 11

The 2006-11 ACLD is a representative sample of almost one million records from the 2006 Census (Wave 1) brought together with corresponding records from the 2011 Census (Wave 2).

This dataset was released in 2013, and updated with additional variables in 2016, including three Visa data items from the Department of Social Services' Settlement Database. It is a rich source for exploring how Australian society has changed between the 2006 and 2011 Censuses.

The original 2006-11 ACLD product has been included for reference. It is recommended for analysis dependent on visa class information.

Linkage Results

3. LINKAGE RESULTS, 2006-2011 (ORIGINAL), 2006 PANEL

At the completion of the linkage process, 800,759 (82%) out of the 979,661 records from the 2006 Census sample (Wave 1) were linked to a 2011 Census (Wave 2) record to create the linked ACLD. This linkage rate was consistent with results from other Bronze linkage projects using the 2006 and 2011 Census.

All results presented in this publication (unless identified in the relevant table) are based on characteristics from the Wave 1 sample and have been confidentialised to prevent the identification of individuals.

Table 2 displays the linkage rate for a range of sub-populations.

TABLE 2 - LINKAGE RATES, By selected characteristics

	2006 Census sample (no.)	ACLD (no.)	Linkage rate (%)
Sex			
Male	480 285	390 487	81.3
Female	499 372	410 274	82.2
Age group (years)			
0-14	194 017	170 834	88.1
15-19	66 247	51 220	77.3
20-24	66 512	49 327	74.2
25-29	62 249	48 642	78.1
30-39	140 271	117 655	83.9
40-49	142 911	123 946	86.7
50-59	126 285	108 962	86.3
60-69	86 385	71 906	83.2
70-74	31 004	23 678	76.4
75 and over	63 784	34 586	54.2
Indigenous status			
Non-Indigenous	942 253	775 419	82.3
Aboriginal	19 697	13 340	67.7
Torres Strait Islander	1 449	923	63.7
Both Aboriginal and Torres Strait Islander	839	543	64.7
Not stated	15 416	10 530	68.3
State/Territory of usual residence			
New South Wales	323 136	263 369	81.5
Victoria	244 095	203 668	83.4
Queensland	192 606	154 013	80.0
South Australia	75 481	62 239	82.5
Western Australia	95 795	77 921	81.3
Tasmania	23 787	19 583	82.3
Northern Territory	8 469	6 226	73.5
Australian Capital Territory	16 186	13 680	84.5
Remote areas			
Major Cities	669 274	552 339	82.5
Inner Regional	195 401	159 611	81.7

Outer Regional	92 396	73 122	79.1
Remote	13 989	10 533	75.3
Very Remote	6 546	4 602	70.3
No Usual Address	2 029	539	26.6
Total(a)(b)(c)	**979 661**	**800 759**	**81.7**

(a) Data presented in the table have been confidentialised. As a result, the sum of individual categories may not align with totals.
(b) Includes Other Territories.
(c) Includes Migratory areas.

The linkage rates that were achieved for the ACLD were relatively consistent across most sub-populations and were in line with expected results. Compared with the national average of 82%, the sub-populations which achieved the highest linkage rates were persons:

- aged 0 to 14 years (88%), followed by 40 to 49 years (87%) and 50 to 59 years (86%)
- of non-Indigenous origin (82%)
- who usually lived in the ACT (85%) and Victoria (83%)
- who usually lived in Major cities (83%).

The subpopulations which achieved the lowest linkage rates were persons:

- aged 20-24 years (74%) and 75 years and over (54%)
- of Aboriginal (68%), Torres Strait Islander (64%) or both Aboriginal and Torres Strait Islander origin (65%)
- who usually lived in the Northern Territory (74%)
- who usually lived in remote (75%) and very remote areas (70%) or who had no usual address in 2006 (27%).

Traditionally, the Census Post Enumeration Survey (PES) has shown that the Census has higher rates of undercount for people of Aboriginal and/or Torres Strait Islander origin, those aged between 20 and 29 and for those in the Northern Territory. As expected, the lower ACLD linkage rates broadly aligned with the same groups that experience higher levels of undercount in the Census. One additional group that had lower linkage rates were persons aged 75 and over at the time of the 2006 Census who, due to age, had an increased risk of death over the ensuing five years. Further information on Census undercount can be found in Census of Population and Housing - Details of Undercount, 2011 (cat. no. 2940.0)

Further data cubes, demonstrating the linkage rates for various sub-populations are available as an attachment to this Information paper.

3.1 LINKAGE ACCURACY

The following quality measures were calculated for the ACLD and indicate a good level of overall quality:

1. The linkage rate, that is the proportion of the 2006 Census sample records linked to a 2011 Census record, including the number of true matches and false links.
2. The consistency of reporting of common information between record pairs.

3.1.1 Linkage Rates, True and False Links

Not all record pairs assigned as links in a data linkage exercise are a match, that is, a record pair belonging to the same individual. While the methodology is designed to ensure that the vast majority of links are true, some are nevertheless false. The linkage strategy used for the ACLD was designed to achieve both a high number of links and to ensure a high level of accuracy to enable longitudinal research. Accordingly, the strategy was restrictive and conservative, especially in the early passes.

Analysis from the results of clerical review was conducted to determine the quality of the linkage process and estimate the number of true links in the linked ACLD file. This process involved calculating the proportion of rejected record pairs at each linkage weight and determining the amount of false links this would represent in the final output file.

Table 3 provides a summary from the results of clerical review, including an estimate of the number of false links accepted in each pass. Due to the nature of deterministic linking and the way in which linked records were retained, no false links were identified in passes 1 and 2. While it is assumed that all links assigned in these passes were true, as they contained consistent information across all key linking fields, in reality there may have been a small but un-quantifiable number of false links.

TABLE 3 - LINKAGE RESULTS, By pass number

		Pass number(a)											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**11**	**12**	**Total(b)**
Links created	(No.)	559 182	131 575	11 131	182 285	212 071	57 713	10 489	10 156	236 180	133 555	29 911	**1 574 248**
Sampled in clerical review	(No.)	30	30	240	400	400	345	206	120	411	201	200	**2 583**
Links assigned	(No.)	544 925	10 919	10 489	62 570	87 248	18 988	1 723	159	50 007	9 827	3 904	**800 759**
Total false links	(No.)	0	0	997	9 929	17 274	1 832	237	29	10 712	1 051	731	**42 792**
False link rate	(%)	0	0	9.5	15.9	19.8	9.6	13.7	18.4	21.4	10.7	18.7	**5.3**

(a) The results of Pass 10 were used to identify the blocking field to be used in Pass 11. As a result, there were no records output from Pass 10.
(b) Data presented in the table have been confidentialised. As a result the sum of individual categories may not align with totals.

The combined clerical review results indicate that the number of false links in the final ACLD file could be as low as 5%. By including a tolerance around these results and assuming a small false link rate for the deterministic passes, the false link rate for the ACLD is estimated to be about 5-10%. The passes that contained the highest proportion of false links were Pass 9 (21.4%), where family information was used to try and resolve unlinked records, and Pass 5 (19.8%), which used a broad geography (SA4) as the blocking field. Whilst this is only an approximate estimate, it does give an indication of the high level of overall quality examined through reviewing a sample of over 2,500 record pairs.

The linkage rate of 82% with a false link rate of 5% was broadly consistent with, or better than, other ABS Census linkage projects which did not use name and address as linkage variables (see Assessing the Likely Quality of the Statistical Longitudinal Census Dataset (cat. no. 1351.0.55.026)).

The conservative and restrictive nature of the blocking and linking strategy helped to minimise the number of estimated false links throughout the linkage process accompanied by quality controls that were implemented during clerical review.

About two-thirds (68%) of all links were achieved in the first pass of the project, which used a deterministic linking methodology to identify and filter matches. In Pass 1, a tight geographic and demographic restriction was implemented to maximise the amount of high quality links assigned and to limit the amount of alternative comparisons required. Using this approach, links were only accepted if a single record pair was identified.

3.1.2 Consistency of Common Information on Record Pairs

In data linkage projects, geographic boundaries function as blocking variables that restrict the search for record pairs. They are also used as linking variables, and when combined with other linking fields such as age, sex and date of birth, provide a high level of uniqueness, and reduce the likelihood of linking to an incorrect record.

Table 4 displays the number of records that had consistent information and is grouped by the consistency of the record pairs across varying levels of geography.

TABLE 4 - CONSISTENCY OF LINKED RECORDS, By geography and selected linking fields

	Consistency of key linkage fields(a)(b)	
	(no.)	**(%)**
Mesh Block combined with		
Age exact, Sex, DOB Day and Month agree	552 714	69.0
Age exact, Sex agree	41 135	5.1
Age +/- 2 years, Sex agree	77 98	1.0
SA2 combined with		
Age +/- 2 years, Sex , DOB Day and Month agree	84 265	10.5
Age +/- 2 years, Sex agree	26 739	3.3
SA4 combined with		
Age +/- 2 years, Sex , DOB Day and Month agree	66 623	8.3
Total records included	779 274	97.3
Total records linked	**800 759**	**100**

(a) Only includes records that agree on all key linking fields.
(b) Categories are mutually exclusive. Records that agree in each category are excluded from subsequent categories.

Just over 97% of all records that were matched in the ACLD linkage process agreed on small to medium levels of geographic area combined with other key linking fields, such as age, sex and date of birth. While the number of consistent fields can give a strong indication of likely linkage quality, other factors should be taken into account, for example, the expected number of people in a geographic area that are likely to share a characteristic by chance. A tolerance of plus or minus two years was used at certain parts of the linkage process to cater for persons who may have understated their age in 2006 and overstated it in 2011 or vice versa.

By contrast, record pairs may have inconsistent information and yet be a true link. Inconsistent information may be recorded for the same person in different Censuses due to a range of factors, including:

- Transcription errors in the Census, where the wrong category is selected or the information is transposed, such as the day the person was born being reported in the month instead of as the day field.
- Data capture errors, where the Census form is scanned using Optical Character Recognition software and certain characters may be mis-classified, such as a 1 captured as a 7 or a 3 as an 8.
- Reporting errors, where information is given for the wrong member of the household (e.g. person 1's information is reported for person 3) or where the person completing the Census estimates information that they do not know (e.g. about a fellow group household member).
- Information that was not stated by the respondent and has been imputed as part of Census processing (such as age or sex).
- A different person fills out the Census form at the different time points and interprets the questions differently.

3.1.2.1 Consistent Reporting of Indigenous Status

Consistency of Indigenous status is a special case, since the change in reporting over time is both a potential indicator of linkage quality, and is of analytical interest.

Results from the 2011 Census observed an unexpected increase in persons who identified as being of Aboriginal and/or Torres Strait Islander origin. This was due, in part, to improvements in Census collection practices that resulted in a more complete enumeration of the Aboriginal and Torres Strait Islander population in 2011 than in 2006. In addition, a significant contributor to this increase, was a change in the propensity of people to identify as being of Aboriginal and/or Torres Strait Islander origin in 2011 compared with 2006 (see Census of Population and Housing: Understanding the Increase in Aboriginal and Torres Strait Islander Counts, 2006-2011 (cat. no. 2077.0)).

While there was a group of people in the ACLD who were identified as non-Indigenous in 2006 and of Aboriginal and/or Torres Strait Islander origin in 2011, this group was relatively small and was counterbalanced by an almost equally sized group who reported the opposite. This pattern of change is different to that expected, given the increasing propensity of people to identify their Aboriginal and Torres Strait Islander origin observed at the aggregate level in the entire 2011 Census.

Throughout the linkage process, Indigenous status was used as a blocking and linking variable. Whilst this would have only made a small contribution to the linkage weight, this may have increased the likelihood of assigning a link to a record pair that contained consistent information for Indigenous status. Record pairs that contained inconsistent information for Indigenous status still had a good chance of being linked, however, providing there was sufficient additional information available for linking.

Differences in the reporting of Indigenous status between 2006 and 2011 on the ACLD may be due to a range of reasons. These include:

- people deliberately identifying their Indigenous origin differently at the two time points
- false links, where similar but not identical persons have been linked
- data capture errors, where multiple boxes may have been selected
- a different person filling out the Census form at each period of time and interpreting the question on Indigenous status differently
- transcription errors in the Census, where the wrong category is selected by accident.

Table 5 shows the reporting of Indigenous status for the linked records on the ACLD, across the 2006 and 2011 Censuses. Further data cubes, demonstrating a more detailed breakdown, by remoteness areas, are provided as an attachment to this Information paper.

TABLE 5 - CONSISTENCY OF INDIGENOUS STATUS FOR LINKED RECORDS, 2006 and 2011

	2011 INDIGENOUS STATUS			
	Non-Indigenous (no.)	**Aboriginal and/or Torres Strait Islander (no.)**	**Not stated (no.)**	**Total (no.)**
2006 INDIGENOUS STATUS				
Non-Indigenous	766 851	1 697	6 868	775 419
Aboriginal and/or Torres Strait Islander	1 367	13 274	165	14 802
Not stated	9 729	226	575	10 530
Total(a)	**777 946**	**15 205**	**7 609**	**800 759**

(a) Data presented in the table have been confidentialised. As a result, the sum of individual categories may not align with totals.

3.2 CHARACTERISTICS OF LINKED AND UNLINKED 2006 CENSUS SAMPLE

Table 6 shows the distribution of key populations across the 2006 Census, the 2006 Census sample and the ACLD.

TABLE 6 - SELECTED CHARACTERISTICS, By 2006 Census, 2006 Census sample and ACLD

					ACLD			
	2006 Census		2006 Census sample		Original results		Weighted results(a)	
	(no.)	(%)	(no.)	(%)	(no.)	(%)	(no.)	(%)
Sex								
Male	9 896 500	49.3	480 285	49.0	390 487	48.8	9 193 092	49.4
Female	10 165 146	50.7	499 372	51.0	410 274	51.2	9 432 201	50.6
State/Territory of usual residence								
New South Wales	6 549 174	32.6	323 136	33.0	263 369	32.9	6 093 946	32.7
Victoria	4 932 422	24.6	244 095	24.9	203 668	25.4	4 624 754	24.8
Queensland	3 904 531	19.5	192 606	19.7	154 013	19.2	3 635 806	19.5
South Australia	1 514 340	7.5	75 481	7.7	62 239	7.8	1 445 720	7.8
Western Australia	1 959 088	9.8	95 795	9.8	77 921	9.7	1 858 559	10.0
Tasmania	476 481	2.4	23 787	2.4	19 583	2.4	465 052	2.5
Northern Territory	192 899	1.0	8 469	0.9	6 226	0.8	179 713	1.0
Australian Capital Territory	324 034	1.6	16 186	1.7	13 680	1.7	319 439	1.7
Age group (years)								
0-9	2 579 496	12.9	127 331	13.0	114 298	14.3	2 551 524	13.7
10-19	2 756 102	13.7	132 937	13.6	107 761	13.5	2 541 650	13.6
20-29	2 684 371	13.4	128 760	13.1	97 973	12.2	2 348 272	12.6
30-39	2 893 058	14.4	140 271	14.3	117 655	14.7	2 800 173	15.0
40-49	2 942 353	14.7	142 911	14.6	123 946	15.5	2 868 511	15.4
50-59	2 574 589	12.8	126 285	12.9	108 962	13.6	2 473 288	13.3
60-69	1 733 297	8.6	86 385	8.8	71 906	9.0	1 640 081	8.8
70-79	1 168 675	5.8	58 277	5.9	42 262	5.3	993 870	5.3
80 and over	729 705	3.6	36 502	3.7	16 002	2.0	408 018	2.2
Indigenous status								
Non-Indigenous	18 266 814	91.1	942 253	96.2	775 419	96.8	17 806 585	95.6
Aboriginal and/or Torres Strait Islander	455 027	2.3	21 985	2.2	14 802	1.8	561 088	3.0
Aboriginal	407 700	2.0	19 697	2.0	13 340	1.7	507 554	2.7
Torres Strait Islander	29 515	0.1	1 449	0.1	923	0.1	32 876	0.2
Both Aboriginal and Torres Strait Islander	17 812	0.1	839	0.1	543	0.1	20 805	0.1
Not stated	1 133 449	5.6	15 416	1.6	10 530	1.3	257 343	1.4
Overseas visitor	206 357	1.0	0	0.0	0	0.0	0	0.0
Total(b)(c)(d)	**20 061 646**	**100**	**979 661**	**100**	**800 759**	**100**	**18 625 246**	**100**

(a) For more information on weighting see chapter 3.4.

(b) Data presented in the table have been confidentialised. As a result, the sum of individual categories may not align with totals.

(c) Includes Other Territories.

(d) Includes Migratory areas.

The distribution of the ACLD file by sub-population was generally well aligned with both the 2006 Census sample and the entire 2006 Census. When looking at the relative difference between these proportions, however, some differences are more clearly observed.

Compared with the entire 2006 Census, the linked ACLD contains relatively more records for people aged 0-9 years, and to a lesser extent those aged 40-49 years, 50-59 years and 60-69 years. By contrast, the ACLD contains relatively fewer records for people aged 20-29 years and 80 years and over. There is also relatively fewer people of Aboriginal and Torres Strait Islander origin in the ACLD, than the entire 2006 Census (1.8% compared with 2.3%). The corresponding weighted estimate, however, represents 3.0% of the total population, which is attributed to benchmarking the 2006 sample to the Aboriginal and Torres Strait Islander population in 2011 and therefore to the higher level of identification observed in the 2011 Census than in 2006 (see section 3.4).

In general, the distribution of weighted counts for the linked ACLD file is close to that of the entire 2006 Census, but it is not designed to produce counts corresponding to the population in 2006. Rather, the weighted population is that of people who were in scope of both the 2006 and 2011 Censuses (see section 3.4). Thus, for example, the lower proportion of older people in the linked file, even after weighting, reflects that impact of deaths on the 2006 sample that occurred between 2006 and 2011.

Further data cubes, demonstrating more detailed population distributions, are provided as an attachment to this Information paper.

3.3 REASONS FOR UNLINKED RECORDS

There are two main reasons why records from the 2006 Census sample were not linked to a 2011 Census record:

1. Records belonging to the same individual were present in the 2006 Census sample and the 2011 Census but these records failed to be linked because they contained missing or inconsistent information.
2. There was no 2011 Census record corresponding to the 2006 Census sample because the person was not counted in the Census.

3.3.1 Missing and/or Inconsistent Information

In these cases, the true match was present in the pool of all record pairs but it was not identified because there was a high level of inconsistency between information on the 2006 Census sample and the 2011 Census record, or key linking fields were missing altogether. The reasons for the match being missed can be categorised into the following groups:

- The missing or inconsistent information did not allow the record pair to be compared in the same blocking categories and could not be linked.
- The record pair did not contain enough common information to distinguish the match from other potential record pairs.
- The record pair was linked, but was attributed a low link weight as it contained a lot of missing or inconsistent information and was positioned below the cut-off identified in sample clerical review.
- The record pair was subjected to clerical review, but the high level of inconsistency did not enable it to be deemed a link.

Accurate address coding was crucial in narrowing the search and differentiating between true and false links. It was a particular challenge for persons who had moved, since linkage was then dependent on the information supplied in 2011 about the person's address in 2006. Processing for the 2011 Census involved coding for address five years ago to a fine level of geography, ideally Mesh block. This was not always possible, either due to the insufficient detail of address information supplied or because by 2011, Census respondents may not have accurately remembered their address on Census Night in 2006.

3.3.2 No 2011 Census Record

A person included in the 2006 Census sample may have had no equivalent 2011 Census record because they were no longer in scope for the Census due to migration from Australia, or death between 2006 and 2011, or they may simply have been missed in the Census.

According to mortality data compiled by the ABS from data supplied by the Registrars of Births, Deaths and Marriages, about 700,000 people died in Australia between 2006 and 2011. If 5% of these people were represented in the 2006 sample, then it could be expected that up to 35,000 people could not have been linked due to death between 2006 and 2011. Similarly, migration data shows that just over one million people left Australia as permanent emigrants over the same period, potentially resulting in up to 50,000 people from 2006 Census sample being unlikely to have a corresponding 2011 Census record.

Due to the size and complexity of the Census, it is inevitable that some people are missed and some are counted more than once. It is for this reason that the Census Post Enumeration Survey (PES) is run shortly after each Census, to provide an independent measure of Census coverage. The PES determines how many people should have been counted in the Census, how many were missed (undercount), and how many were counted more than once (overcount). It also provides information on the characteristics of those in the population who have been missed or overcounted.

The net undercount rate for the 2011 Census was 1.7%, with a higher rate for Aboriginal and Torres Strait Islander people than for the non-Indigenous population (see Census of Population and Housing - Details of Undercount, 2011 (cat. no. 2940.0)) Thus, roughly, 15,000 people from the 2006 Census sample could have been missed in the 2011 Census. This estimate is a starting point only and does not take into account the likelihood of people being missed in successive Censuses.

When taking into account all of these factors, it is estimated that over half of the unlinked 2006 Census sample (100,000 out of the 180,000 unlinked records) would not have a corresponding record in the 2011 Census. This would indicate that the initial linkage rate of 82% could be representative of up to 91% of the population that actually had an opportunity to be linked.

3.4 WEIGHTING

Weighting is the process of adjusting a sample to infer results for the relevant population. To do this, a 'weight' is allocated to each sample unit - in this case, persons. The weight can be considered an indication of how many people in the relevant population are represented by each person in the sample. Weights were created for linked records in the ACLD to enable longitudinal population estimates to be produced. Cross-sectional population estimates for 2006 and 2011 are available from each Census.

The ACLD began as a random sample of 5% of the Australian population in 2006. As such, each person in the sample should represent about 20 people in the population. Between Censuses, however, the in scope population changes as people die or move overseas. In addition, Census net undercount and data quality can affect the capacity to link equivalent records across waves. The ACLD weighting process, benchmarked the linked ACLD records to the population that was in scope of both the 2006 and 2011 Censuses. The weights were based on four components: the design weight, undercoverage adjustment, missed link adjustment and population benchmarking.

The original population benchmark was the 2011 Estimated Resident Population (ERP). The 2011 ERP was chosen over the 2006 ERP as the baseline population as it is more recent. The ERP was than adjusted to exclude births and overseas arrivals that had occurred between 2006 and 2011.

Weights were benchmarked to the following population groups:

- state by age (ten year groups), by sex, by mobility (interstate arrivals benchmarked separately)
- Indigenous status by state.

The weights have a mean value of 24 and range between 17 and 103. Higher weights are associated with people of Aboriginal and Torres Strait Islander origin and people who moved interstate between 2006 and 2011. For more information see the Appendix.

About this Release

The Australian Census Longitudinal Dataset (ACLD) uses data from the Census of Population and Housing to build a rich longitudinal picture of Australian society. The ACLD can uncover new insights into the dynamics and transitions that drive social and economic change over time, and how these vary for diverse population groups and geographies.

The ACLD is a random five per cent sample of the Australian population and three waves of data have so far contributed to the ACLD from the 2006 Census, 2011 Census and 2016 Census.

The 2006-2016 ACLD is produced from a representative sample of records from the 2006 Census was brought together with corresponding records from the 2011 and 2016 Censuses to form the 2006 Panel of the ACLD. Additionally, records from the 2011 Census were brought together with corresponding records from the 2016 Census to form the 2011 Panel of the ACLD. While the 2011 Panel significantly overlaps with the 2006 Panel, the 2011 Panel includes new births and migrants since the 2006 Census.

The 2006-2016 ACLD is a rich source for exploring how Australian society has changed between the 2006, 2011 and 2016 Censuses.

This release provides in depth information about the sampling and linking methodologies, and linkage results.

History of Changes

HISTORY OF CHANGES

20/03/2019

Information added relating to the 2006-11-16 linkage.

Formatting changes to combine the two separate publications, *Information Paper: Australian Census Longitudinal Dataset, Methodology and Quality Assessment, 2006-2011* and *Information Paper: Australian Census Longitudinal Dataset, Methodology and Quality Assessment 2011-2016,* into one publication.

Explanatory Notes

References

REFERENCES

Australian Bureau of Statistics:
Assessing the Likely Quality of the Statistical Longitudinal Census Dataset, (2009) cat. no 1351.0.55.026.
Australian Census Longitudinal Dataset: Methodology and Quality Assessment, 2006-2011, Information Paper (2013) cat. no. 2080.5.
Australian Census Longitudinal Dataset with Social Security and Related Information, experimental statistics, 2006-2011, Microdata (2017) cat. no. 2085.0.
Australian Census and Migrants Integrated Dataset, 2011, (2014) cat. no. 3417.0.55.001
Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, July 2016, (2016) cat. no. 1270.0.55.001.
Census Data Enhancement: An Update (2010) cat. no. 2062.0.
Census of Population and Housing: Understanding the Increase in Aboriginal and Torres Strait Islander Counts, 2006-2011, (2013) cat. no. 2077.0.
Census of Population and Housing - Details of Overcount and Undercount, Australia 2016, (2016) cat. no. 2940.0.
Census of Population and Housing - Details of Undercount, 2011, (2012) cat. no. 2940.0
Deaths, Australia, 2016, cat. no. 3302.0.
Deaths, Australia, 2015, cat. no. 3302.0.
Deaths, Australia, 2014, cat. no. 3302.0.
Deaths, Australia, 2013, cat. no. 3302.0.
Deaths, Australia, 2012, cat. no. 3302.0.
Deaths, Australia, 2011, cat. no. 3302.0.
Death registrations to Census linkage project - Key Findings for Aboriginal and Torres Strait Islander peoples, 2011-12, (2013) cat. no.3302.0.55.005.
Life tables for Aboriginal and Torres Strait Islander Australians, 2010-2012, (2013) cat. no. 3302.0.55.003
Migration, Australia, 2015-16, (2017) cat. no. 3412.0.
Migration, Australia, 2014-15, (2016) cat. no. 3412.0.
Migration, Australia, 2013-14, (2015) cat. no. 3412.0.
Migration, Australia, 2011-12 and 2012-13, (2013) cat. no 3412.0.
Statistical Longitudinal Census Dataset, 2006-2011, (2013) cat. no. 2080.0.
Understanding the Census and Census Data, Australia, 2016, cat. no. 2900.0.
A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset, August 2009, Research Paper (2009) cat. no. 1351.0.55.025.

Australian Institute of Health and Welfare and Australian Bureau of Statistics (2012). "National best practice guidelines for data linkage activities relating to Aboriginal and Torres Strait Islander people", cat. no. IHW 74. AIHW, Canberra.

Chipperfield, J., Hansen, N. and Rossiter, P. (2018) "Estimating Precision and Recall for Deterministic and Probabilistic Record Linkage*", International Statistical Review.*

Chipperfield, J., Brown, J. and Watson, N. (2017) "The Australian Census Longitudinal Dataset: using record linkage to create a longitudinal sample from a series of cross-sections", *Australian & New Zealand Journal of Statistics*, 59(1), pp. 1-16.

Christen, P. and Churches, T. (2005) "Febrl 0.3 Documentation".

Christen, P., Churches, T. and Hegland, M. (2004) “Febrl – A Parallel Open Source Data Linkage System”, Proceedings of the 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, pp. 638-647.

Conn, L. and Bishop, G. (2006) “Exploring Methods for Creating a Longitudinal Census Data Set”, Methodology Advisory Committee Papers, cat. no. 1352.0.55.076, Australian Bureau of Statistics, Canberra.

Cross Portfolio Statistical Integration Committee (2010), "High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes", CPSIC, Canberra.

Fellegi, I. and Sunter, A. (1969) “A Theory for Record Linkage”, *Journal of the American Statistical Association*, 64(328), pp. 1183–1210.

Fellegi, Ivan P. and Sunter, Alan B. (1969) “A Theory for Record Linkage”, Journal of the American Statistical Association, 64(328), pp. 1183–1210.

Harding, S., Jackson Pulver, L., McDonald, P., Morrison, P., Trewin, D. and Voss, A. (2017). Report on the quality of 2016 Census data.

Samuels, C. (2012) “Using the EM Algorithm to Estimate the Parameters of the Fellegi–Sunter Model for Data Linking”, Methodology Advisory Committee Papers, cat. no. 1352.0.55.120, Australian Bureau of Statistics, Canberra.

Appendix: Weighting the 2011-16 ACLD (Appendix)

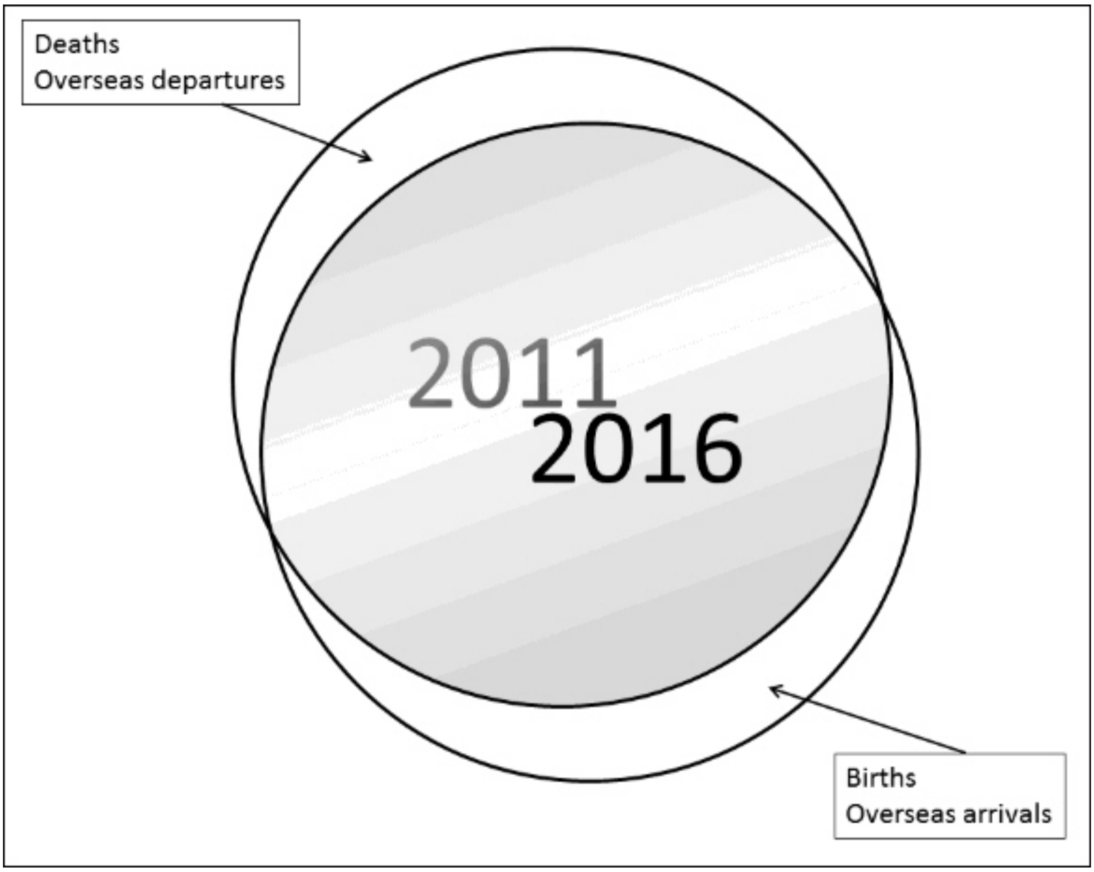
APPENDIX: WEIGHTING THE ACLD

INTRODUCTION

The process of weighting enables the data user to estimate the number of people in the whole population with particular characteristics based on the observations from a sample. To do this, a 'weight' is allocated to each sample unit. The value of the weight indicates how many population units are represented by the sample unit.

The ACLD is designed to measure change in Australian society over time. For the 2011 ACLD Panel, a longitudinal weight has been implemented which allows the weighted sample to represent all persons who were in scope of both of the 2011 and 2016 Censuses. As shown in Figure 1, this ‘longitudinal population’ is the overlap between the two Censuses (the shaded region). To estimate the size of this population, the 2016 Estimated Resident Population (ERP) was multiplied by the estimated probability that a person was in scope in 2011, calculated using the 2016 Census responses for the reported 5 year ago address. Further information on this approach can be found in the paper Chipperfield, Brown & Watson (2017).

FIGURE 1 - IN SCOPE POPULATION FOR THE AUSTRALIAN CENSUS LONGITUDINAL DATASET, 2011-2016



This method for estimating the relevant overlapping 2011 and 2016 populations is an improvement over the method used in the original linkage of the 2006-2011 ACLD. The current approach overcomes the limitations of the previous approach which inaccurately accounted for:

- people who were overseas arrivals after one Census and who subsequently left Australia (or died) before the following Census; and
- people who left Australia after one Census and then returned to Australia before the following Census.

CALCULATING LONGITUDINAL WEIGHTS

Longitudinal weights were calculated for each 2011 Panel sample record that was linked to a 2016 Census record. No weights were calculated for the unlinked records. The longitudinal weight for a linked sample record in this release of the ACLD is a measure of how many people it represents in the 2011 and 2016 overlapping populations. The weights consist of three components. The first component reflects the probability of a record being selected in the 2011 Panel sample. The second component takes into account that some selected records are less likely to be linked than other selected records. The third component takes into account Census undercount (e.g. a person in an undercounted population is less likely to have a Census record and so is less likely to have a selected Census record) and ensures that the weights are consistent with population benchmarks. These three components lead to the final weight, calculated as:

Final weight = (Design weight) x (Missed link adjustment) x (Calibration to known population totals adjustment).

Design Weight

The records in the 2011 ACLD Panel sample were selected from the 2011 Census population using equal probability random selection. For a sample size of 5.7%, the design weight for all records of the ACLD is the inverse of the probability of selection, and is equal to 17.5.

Missed Link Adjustment

A missed link occurs when a 2011 sampled record has a corresponding 2016 Census record, but the link is not identified. As missed links are more likely to occur in certain population groups, not making this adjustment would mean that these population groups would be under-represented in the linked sample. The missed link adjustment weight is equal to the inverse of the estimated propensity to link. No attempt was made to correct for false links.

The propensity to link was estimated using a logistic regression model that was applied to the 2011 Panel sample with link status as the response variable. The logistic regression model describes a relationship between a 2011 sample record's propensity to link and its values for a range of 2011 Census variables such as Indigenous status, marital status, country of birth, language spoken at home and English proficiency, labour force participation and occupation, educational attainment, mobility (whether moved address in the preceding year) and remoteness. It was found that the estimated propensity to link varied considerably between records.

Two separate models were applied to the 2011 Panel sample. The first model was applied to people under the age of 15 years on 2011 Census night. This model excluded the variables that were not applicable to people under 15 years of age, such as marital status and labour force participation. The second model was applied to the remainder of the sample (persons aged 15 years or over in 2011).

The missed link adjustment carries the assumption that the ACLD contains no false links, while not assuming that all records in the 2011 sample have a corresponding 2016 Census record.

Odds ratios and accompanying Wald confidence intervals for the predictor variables for the first model (for persons aged under 15 years in 2011) are contained in Table A.1. A comparison group is selected for each characteristic, and the odds ratio for the other categories represents the ratio of the odds of being linked in contrast to the comparison group. For instance, Table A.1 shows the odds ratios by age group in 2011. Those persons aged under 15 years with English proficiency ‘Not at All’ were less likely to be linked than those with English proficiency ‘Very Well’ (the comparison group), but more likely than those with English proficiency ‘Not Well’. Conversely, the odds ratios for remoteness in 2011 show that persons aged under 15 years reporting 'Inner Regional' were more likely to be linked than those reporting 'Major City' (the comparison group).

TABLE A.1 - ODDS RATIOS FROM THE LOGISTIC REGRESSION MODEL, Persons aged under 15 years, 2011

		95% CONFIDENCE LIMITS	
Selected characteristics	**Odds ratio**	**Low limit**	**Upper limit**
AGE GROUP			
0-7 years (comparison group)	1.000	. .	. .
8-13 years	1.013	1.000	1.026
14 years	1.011	0.989	1.033
COUNTRY OF BIRTH			
Oceania and Antarctica (non-Indigenous persons) (comparison group)	1.000	. .	. .
Americas	0.577	0.525	0.635
Indigenous Australian	0.872	0.847	0.897
North Africa & Middle East	0.465	0.423	0.510
North-East Asia	0.526	0.483	0.572
North-West Europe	0.661	0.630	0.693
South-East Asia	0.569	0.533	0.606
Southern & Central Asia	0.620	0.582	0.660
Southern & Eastern Europe	0.425	0.360	0.501
Sub Saharan Africa	0.722	0.668	0.780
Missing(a)	0.773	0.710	0.841

LANGUAGE			
English (comparison group)	1.000	. .	. .
Australian Indigenous	1.087	0.976	1.211
East Asian	0.992	0.954	1.032
East European	0.973	0.916	1.034
Other North European	0.967	0.891	1.050
South Asian	1.004	0.955	1.056
South East Asian	0.987	0.943	1.033
South European	0.977	0.934	1.021
South West and Central Asian	0.984	0.940	1.029
Other	0.805	0.754	0.859
Missing(a)	0.882	0.803	0.969
ENGLISH PROFICIENCY			
Very Well (comparison group)	1.000	. .	. .
Well	0.777	0.743	0.812
Not Well	0.644	0.589	0.704
Not at All	0.806	0.658	0.987
Missing	0.939	0.841	1.049
MOBILITY			
Same usual address one year ago (comparison group)	1.000	. .	. .
Different usual address one year ago	0.838	0.829	0.848
SCHOOL SECTOR			
Government (comparison group)	1.000	. .	. .
Catholic	1.042	1.026	1.058
Other Non-Government	1.057	1.039	1.076
Missing/Other(b)	0.874	0.853	0.896
REMOTENESS			
Major City (comparison group)	1.000	. .	. .
Inner Regional	1.021	1.006	1.036
Outer Regional	1.010	0.989	1.031
Remote	0.961	0.913	1.012
Very Remote	0.936	0.866	1.011
Other(c)	1.047	0.813	1.348

. . Not applicable
(a) Includes Supplementary codes
(b) Includes other school sector and pre-school
(c) Includes Migratory, Offshore and Shipping Zones and No usual address
Source: Australian Census Longitudinal Dataset

Odds ratios and accompanying Wald confidence intervals for the predictor variables in the second model (for persons aged 15 years or over in 2011) are contained in Table A.2. A wider variety of variables were available for this age group. In comparison to those persons aged under 15, persons aged 15 years or over with English proficiency ‘Not Well’ were less likely to be linked than those with English proficiency ‘Very Well’ (the comparison group), but more likely than those with English proficiency ‘Not At All’. Conversely, the odds ratios for age in 2011 show that persons reporting aged 65 and over were more likely to be linked than those reporting age 15-24 (the comparison group). This was similar for persons reporting 'Clerical and Administrative' and 'Sales Workers' for Occupation as they were more likely to be linked than those that were 'Not in the labour force' (the comparison group).

TABLE A.2 - ODDS RATIOS FROM LOGISTIC REGRESSION MODEL, persons aged 15 years or over, 2011

		95% CONFIDENCE LIMITS	
Selected characteristics	**Odds ratio**	**Low limit**	**Upper limit**
AGE GROUP			
15-24 years (comparison group)	1.000	. .	. .
25-34 vs 15-24	0.841	0.832	0.850
35-44 vs 15-24	0.968	0.957	0.978
45-54 vs 15-24	1.008	0.997	1.020
55-64 vs 15-24	1.031	1.019	1.043
65+ vs 15-24	1.080	1.067	1.094
COUNTRY OF BIRTH			
Oceania and Antarctica (non-Indigenous persons) (comparison group)	1.000	. .	. .
Americas	0.829	0.810	0.848
Indigenous Australian	0.955	0.939	0.971
North Africa & Middle East	0.951	0.928	0.976
North-East Asia	0.748	0.730	0.766
North-West Europe	0.951	0.943	0.959
South-East Asia	0.894	0.877	0.911
Southern & Central Asia	0.838	0.820	0.858
Southern & Eastern Europe	0.976	0.961	0.992
Sub Saharan Africa	0.930	0.912	0.949
Missing	0.884	0.867	0.900
LANGUAGE			
English (comparison group)	1.000	. .	. .
Australian Indigenous	0.978	0.925	1.035
East Asian	1.004	0.983	1.026
East European	1.006	0.986	1.026
Other North European	0.926	0.902	0.950
South Asian	0.947	0.922	0.973
South East Asian	0.969	0.948	0.990
South European	0.953	0.940	0.967
South West and Central Asian	0.989	0.965	1.013
Missing	0.951	0.918	0.985
Other	0.836	0.813	0.860
ENGLISH PROFICIENCY			
Very Well (comparison group)	1.000	. .	. .
Well	0.859	0.848	0.871
Not Well	0.826	0.811	0.840
Not at All	0.661	0.636	0.687
Not stated	0.950	0.912	0.990
MOBILITY			
Same usual address one year ago (comparison group)	1.000	. .	. .
Different usual address one year ago	0.797	0.793	0.801
Remoteness			
Major City (comparison group)	1.000	. .	. .
Inner Regional	1.005	0.999	1.011
Outer Regional	0.979	0.971	0.987
Remote	0.938	0.918	0.959
Very Remote	0.845	0.818	0.874
Other(b)	0.948	0.894	1.005
REGISTERED MARITAL STATUS			
Married (comparison group)	1.000	. .	. .
Separated	0.951	0.938	0.964
Divorced	0.939	0.931	0.947
Widowed	0.991	0.980	1.002

Never Married	0.940	0.933	0.946
HIGHEST YEAR OF SCHOOL COMPLETED			
Year 12 (comparison group)	1.000	. .	. .
Year 11	1.010	1.002	1.018
Year 10	0.996	0.989	1.003
Year 9 or below(a)	1.003	0.995	1.012
Not stated	0.990	0.974	1.007
LABOUR FORCE STATUS AND OCCUPATION			
Not in the labour force (comparison group)	1.000	. .	. .
Unemployed	0.954	0.942	0.966
Employed			
Professional	1.080	1.070	1.090
Manager	1.045	1.035	1.055
Technicians and trades	1.038	1.028	1.049
Community and personal service	1.056	1.045	1.067
Clerical and administrative	1.076	1.065	1.086
Sales workers	1.075	1.063	1.086
Machinery operators and drivers	1.016	1.003	1.030
Labourers	0.999	0.988	1.010
Employed, occupation not stated	0.865	0.843	0.887
Not stated	0.889	0.863	0.916
LEVEL OF NON-SCHOOL QUALIFICATION			
No post-school qualification (comparison group)	1.000	. .	. .
Postgraduate Degree	0.969	0.958	0.982
Graduate Diploma and Graduate Certificate	1.028	1.012	1.045
Bachelor Degree	0.985	0.977	0.993
Advanced Diploma and Diploma	1.009	1.001	1.018
Certificate	1.021	1.015	1.028
Level of non-school qualification not stated or inadequately described	0.948	0.936	0.960

. . Not applicable
(a) Includes persons who did not go to school.
(b) Includes Migratory, Offshore and Shipping Zones and No usual address
Source: Australian Census Longitudinal Dataset, 2011-2016

Calibration to Longitudinal Population Totals

At this point in the process an intermediate weight had been calculated for each linked sample record that was equal to (Design weight) x (Missed link adjustment). This intermediate weight was then calibrated (or adjusted) so that the resulting weighted counts of the ACLD links would be equal to estimates of the longitudinal population size at the national and selected sub-national levels. The two sets of longitudinal population groups calibrated to were:

1. state/territory, by sex, by ten year age group;
2. Indigenous status by state/territory.

The size of these longitudinal population groups was estimated by multiplying the 2016 ERP for each group by the estimated proportion of 2016 Census responders for that group who reported being in scope of the 2011 Census, i.e. resident at an Australian address on 2011 Census night. These proportions were estimated for the cross-classification of the state/territory, age group, and sex (Table A.3) and separately for the cross-classification of state and Indigenous status (Table A.4) using the responses to the 2016 Census address five years ago question. For example, Table A.3 indicates that 96.7% of males in NSW aged 45-54 reported being in scope of the 2011 Census.

TABLE A.3. ESTIMATED PROPORTION OF 2016 CENSUS RESPONDENTS IN SCOPE OF THE 2011 CENSUS, By state/territory, sex and age

STATE/TERRITORY 2011 AND 2016 **Age group (years)**	**Males** **(no.)**	**Females** **(no.)**
NEW SOUTH WALES		
5-14	0.949	0.949
15-24	0.889	0.883
25-34	0.835	0.823
35-44	0.918	0.925
45-54	0.967	0.967
55-64	0.979	0.974
65-74	0.984	0.983
75-84	0.991	0.991
85 or over	0.995	0.995
VICTORIA		
5-14	0.938	0.938
15-24	0.864	0.858
25-34	0.842	0.824
35-44	0.917	0.921
45-54	0.964	0.965
55-64	0.978	0.973
65-74	0.984	0.984
75-84	0.992	0.992
85 or over	0.996	0.996
QUEENSLAND		
5-14	0.948	0.948
15-24	0.920	0.910
25-34	0.881	0.864
35-44	0.937	0.935
45-54	0.968	0.968
55-64	0.980	0.979
65-74	0.985	0.986
75-84	0.992	0.992
85 or over	0.995	0.995
SOUTH AUSTRALIA		
5-14	0.944	0.946
15-24	0.910	0.909
25-34	0.896	0.876
35-44	0.930	0.932
45-54	0.973	0.975
55-64	0.987	0.984
65-74	0.990	0.991
75-84	0.995	0.995
85 or over	0.998	0.998
WESTERN AUSTRALIA		
5-14	0.915	0.916
15-24	0.893	0.893
25-34	0.840	0.817
35-44	0.895	0.898
45-54	0.948	0.952
55-64	0.975	0.972
65-74	0.983	0.983
75-84	0.991	0.990
85 or over	0.996	0.995
TASMANIA		
5-14	0.976	0.974

15-24	0.952	0.949
25-34	0.928	0.922
35-44	0.965	0.965
45-54	0.985	0.985
55-64	0.992	0.993
65-74	0.993	0.995
75-84	0.996	0.996
85 or over	0.998	0.997
NORTHERN TERRITORY		
5-14	0.941	0.939
15-24	0.928	0.920
25-34	0.873	0.842
35-44	0.912	0.917
45-54	0.955	0.962
55-64	0.972	0.970
65-74	0.981	0.985
75-84	0.987	0.985
85 or over	0.984	0.993
AUSTRALIAN CAPITAL TERRITORY		
5-14	0.922	0.921
15-24	0.862	0.853
25-34	0.867	0.844
35-44	0.906	0.909
45-54	0.951	0.956
55-64	0.969	0.966
65-74	0.982	0.982
75-84	0.991	0.990
85 or over	0.998	0.997

Source: 2016 Estimated Resident Population

TABLE A.4. ESTIMATED PROPORTION OF 2016 CENSUS RESPONDENTS IN SCOPE OF THE 2011 CENSUS, By state/territory and Indigenous status

State/Territory	**Aboriginal and Torres Strait Islander persons (no.)**	**Other persons(a) (no.)**
New South Wales	0.996	0.927
Victoria	0.993	0.922
Queensland	0.996	0.941
South Australia	0.997	0.947
Western Australia	0.997	0.914
Tasmania	0.997	0.971
Northern Territory	0.999	0.894
Australian Capital Territory	0.988	0.914

(a) Includes non-Indigenous persons and persons who did not state an Indigenous status in 2016.
Source: 2016 Estimated Resident Population

At this point an intermediate weight was calculated for each linked sample record that was equal to (Design weight) x (Missed link adjustment). This intermediate weight was calibrated (or adjusted) so that the resulting weighted counts of the ACLD links would be equal to the estimated longitudinal population sizes in Table A.3 and A.4. The intermediate weights were calibrated using a 'raking' tool. This is a program which was developed to determine record level weights using iterative horizontal and vertical passes through the unit records until a satisfactory set of final weights are converged upon. Imposing bounds on the calibration adjustment was not necessary because extremely high or low final weights were not produced.

This calibration adjustment improves the accuracy of weighted estimates and it implicitly adjusts for the small proportion of people who were in scope for the 2011 Census but who did not complete a Census form in 2016.

Summary of weights

The mean weight for selected characteristics gives an indication of how much the final weight differs from the initial design weight (17.5) in order to address missed links and Census undercount. Table A.5 shows that the mean final weight for the linked records is 22.3 for females, and 23.2 for males. The largest weight was 83 and the smallest was 14.8. The mean weight was higher for Aboriginal and Torres Strait Islander persons (30.4) and for people in the Northern Territory (30.2).

TABLE A.5 - DESCRIPTIVE STATISTICS FOR WEIGHTS, by Selected Characteristics, 2016

	Count (a)	Minimum Weight	Maximum Weight	Mean Weight	Standard Deviation	Median Weight
SEX						
Male	450 054	14.8	83.1	23.2	4.9	22.0
Female	477 460	14.8	81.9	22.3	4.7	21.3
AGE						
0-14	126 917	16.1	81.9	22.3	5.1	21.0
15-24	119 827	16.5	67.5	23.5	4.0	22.7
25-34	109 438	17.4	83.1	27.9	5.3	27.3
35-44	127 858	15.9	60.6	23.3	4.7	22.7
45-54	139 010	16.0	68.8	21.9	4.1	20.8
55-64	127 797	15.6	62.7	21.3	3.8	20.1
65-74	99 605	15.2	55.5	20.6	3.4	19.4
75-84	54 414	14.8	58.0	20.2	3.5	18.8
85 or over	22 651	14.8	56.2	21.2	3.9	19.8
INDIGENOUS STATUS						
Aboriginal and/or Torres Strait Islander	23 059	18.9	83.1	30.4	6.0	29.9
Other (b)	904 457	14.8	81.9	22.5	4.6	21.5
STATE/TERRITORY OF USUAL RESIDENCE						
New South Wales	296 695	15.5	81.9	22.7	4.9	21.6
Victoria	234 629	15.7	80.5	22.7	4.7	21.6
Queensland	185 576	15.7	81.2	23.0	4.6	22.2
South Australia	71 025	14.8	68.7	21.5	4.3	20.3
Western Australia	95 198	15.4	73.8	23.0	5.2	21.7
Tasmania	21 781	15.9	58.8	21.8	3.8	20.7
Northern Territory	6 929	17.1	83.1	30.2	8.0	28.7
Australian Capital Territory	15 583	15.2	60.0	22.0	4.6	21.1

a) Counts presented in the table have been perturbed.
b) Includes non-Indigenous persons and persons who did not state an Indigenous status in 2016.
Source: ABS, Australian Census Longitudinal Dataset.

Quality Declaration

QUALITY DECLARATION

INSTITUTIONAL ENVIRONMENT

For information on the institutional environment of the Australian Bureau of Statistics (ABS), including the legislative obligations of the ABS, financing and governance arrangements, and mechanisms for scrutiny of ABS operations, see ABS Institutional Environment.

In April 2012, the ABS became an accredited Integrating Authority under the Commonwealth data integration interim arrangements. A copy of the accreditation claims made by the ABS, which have been verified by an independent auditor, is available on data.gov.au. The ABS only undertakes data integration for statistical and research purposes and where there is a strong public benefit in doing so.

The Australian Census Longitudinal Dataset (ACLD) is released in TableBuilder and as a microdata product in the DataLab. Microdata files are released in accordance with the conditions specified in the Statistics Determination section of the Census and Statistics Act 1905. This ensures that confidentiality is maintained whilst enabling micro level data to be released. More information on the confidentiality practices associated with TableBuilder can be found in TableBuilder, User Guide (cat. no. 1406.0.55.005) on the Confidentiality page. To protect confidentiality of data within the DataLab, users are supervised at all times and must not bring mobile phones, cameras, USB keys, laptops, palm pilots or similar transmission or storage devices into the secure location. All outputs produced by users in DataLab are manually cleared for release after the session.

RELEVANCE

Data for the Census of Population and Housing used in this product were collected on 8 August 2006, 9 August 2011 and 9 August 2016. The scope of the Census is all persons enumerated in Australia on Census night. The Census covers all areas in Australia and includes persons living in both private and non-private dwellings but excludes:

- diplomatic personnel of overseas governments and their families
- Australian residents overseas on Census Night and

The ACLD is built upon a 5% sample of records taken from a particular Census that is then linked to following Censuses. There are currently two samples, 2006 and 2011, with each being representative of the Australian population at the time of the Census collection.

Overseas visitors are excluded from the 2006 and 2011 ACLD Panel samples. Visitors within Australia to private and non-private dwellings on Census Night are included

The Census collects information on demographics, income, labour force, unpaid work, dwelling characteristics and family and household relationships.

For more information, see How Australia Takes a Census, 2006 (cat. no. 2903.0), How Australia Takes a Census, 2011 (cat. no. 2903.0), Census of Population and Housing: Understanding the Census and Census Data, Australia, 2016 (cat. no. 2900.0), and the 2006, 2011 and 2016 issues of the Census Dictionary (cat. no. 2901.0).

TIMELINESS

The Census of Population and Housing is conducted every five years. For further information see the publications How Australia Takes a Census, 2006 (cat. no. 2903.0), How Australia Takes a Census, 2011 (cat. no. 2903.0) and Census of Population and Housing: Understanding the Census and Census Data, Australia, 2016 (cat. no. 2900.0).

The first wave of Census data for the ACLD was from 2006, the second wave was from 2011, and the third wave was from 2016.

Microdata from the 2006-11 ACLD was first made available in December 2013. The 2011-16 data was available from February 2018, and the 2006-11-16 data available from March 2019.

ACCURACY

The ACLD is a random 5% sample of persons enumerated in Australia on either Census Night, 2006 or Census Night, 2011 which has been linked using statistical techniques to records from successive Censuses. False links can occur during the linkage process as even when a record pair matches on all or most linking fields, it may not actually belong to the same individual. The nature of the process used for the ACLD linkage means that while the methodology is designed to ensure links obtained are to a high degree of accuracy, some false links may be present within the ACLD dataset. There is an estimated 5-10% false link rate in the original linkage of the 2006-2011 ACLD, an estimated 5% false link rate in the re-link of the 2006-2011 ACLD and an estimated 1% false link rate in the 2011-2016 linkages.

Sampling error occurs because only a small proportion of the total population is used to produce estimates that represent the whole population. Sampling error refers to the fact that for a given sample size, each sample will produce different results, which will usually not be equal to the population value. There are two common ways of reducing sampling error - increasing sample size and/or utilising an appropriate selection method (for example, multi-stage sampling would be appropriate for household surveys). Given the large sample size for the ACLD (1 in 20 persons), and simple random selection, sampling error is minimal.

The ACLD sample was weighted to an estimate of the population that was resident in Australia on Census Night for the relevant linkage periods. For example, the linkage of the 2011 Panel to the 2016 Census is weighted to an estimate of the population that was resident in Australia at both the 2011 and 2016 Censuses. The weights adjust for missed links and Census undercount.

Information on methodology, linkage quality and weighting can be found in Information Paper: Australian Census Longitudinal Dataset, Methodology and Quality Assessment, ACLD (cat. no. 2080.5). Steps are taken to confidentialise the data made available on TableBuilder in such a way as to maximise the usefulness of the content while maintaining the confidentiality of respondents selected in the ACLD sample. As a result it may not be possible to exactly reconcile all the statistics produced from the microdata with other published statistics. Further information about the steps taken to confidentialise the microdata can be found in TableBuilder, User Guide (cat. no. 1406.0.55.005) on the Confidentiality page.

COHERENCE

A small percentage of linked records have inconsistent data, such as a different country of birth at the two time points or an age inconsistency of more than one year. Inconsistencies may be due to:

- false link - the record pair does not belong to the same individual
- reporting error - information for the same individual was reported differently in 2011 and in 2016
- processing error - the value of a data item was inaccurately assigned or imputed during processing.

ACLD microdata contains a large number of data items and in some cases the level of detail has been collapsed from that described in the Census Dictionary. For more information on the level of detail provided, please see the associated Data Items list.

While the 2011 and 2016 Censuses had predominantly the same questions and were processed in a similar way, there were some differences between them.

For example, a number of changes were made to how industry of employment information was collected for the 2016 Census. The ABS advises this data is not directly comparable to the previous Census Industry of employment data, and should not be used to measure longitudinal transitions between industries from 2011 to 2016. For further information refer to Industry of Employment (INDP) in Census of Population and Housing: Understanding the Census and Census Data, Australia, 2016 (cat. no. 2900.0).

Notable data items that are different between Census years are personal, family and household income. Income was collected in ranges and these ranges are different in different Census years. The ACLD does not include an adjustment to income data for inflation.

Some data items were derived differently between Censuses. In these instances, to aid comparability, the 2006 and 2011 variables were re-derived to make them consistent with the 2016 derivation.

For more information on the differences between the 2006, 2011 and 2016 Censuses see What's New for 2011? and What's New for 2016?

Estimates derived from the ACLD may differ to those derived from other sources. This is due to a range of factors including:

- **Collection methodology**
  The ACLD is derived from Census data that is self-reported by households across Australia on Census night. This will differ from other ABS collections which may rely on different collection methodologies (e.g. trained interviewers, administrative sources). In addition, the way survey questions are phrased and the answer options available for a given question may affect the information provided by respondents.

- **Reference period**
  The reference periods for the ACLD are the Census nights of each year. Other collections may use different reference periods.

- **Sampling design**
  The ACLD uses a 5% sample of Census data as its base population. This will differ from other collections that may collect information from the entire population of Australia (e.g. the Census) or from a sample of dwellings (e.g. Labour Force Survey).

- **Sampling and non-sampling error**
  While every effort is made to minimise error, each collection will have some level of error. Survey collections are subject to some level of sampling error, as they are based on information obtained from a sample of dwellings or businesses. The Census is not subject to this type of error, but is subject to some level of undercount. The ACLD is constructed using a sample of records from the Census, and is therefore subject to a level of sampling error of its own.

- **Scope and coverage**
  The ACLD weights benchmark the linked records to the longitudinal population that was in scope of both or all three Censuses. This will be different to cross-sectional estimates which may be benchmarked to a point-in-time population, such as the Estimated Resident Population.

- **Linkage error**
  The ACLD is subject to linkage error, as records from one Census are linked to corresponding records from the subsequent Census. While every effort is made to minimise false links, they can occur. Linkage error will not be apparent in other collections which are

not produced through data integration.

For these reasons, while the results from the ACLD are considered to be broadly representative of the Australian population, they are not strictly comparable with statistics derived from other collections.

For detailed information about the different methodologies for each collection, refer to the Explanatory Notes within each release.

For detailed information regarding the differences between the Census and Labour Force collections, refer to The 2016 Census and the Labour Force Survey in Census of Population and Housing: Understanding the Census and Census Data, Australia, 2016 (cat. no. 2900.0).

For detailed information regarding Census data, including changes to Census questions and data quality statements for each Census data item, refer to Understanding the data in Census of Population and Housing: Understanding the Census and Census Data, Australia, 2016 (cat. no. 2900.0).

INTERPRETABILITY

This publication should be referred to when using the microdata. It contains information on the Methodology, File Structure, Using the ACLD in TableBuilder, The ACLD in the DataLab, Conditions of Use and the Data Items list.

Detailed information on methodology, linkage quality and weighting can be found in Information Paper: Australian Census Longitudinal Dataset, Methodology and Quality Assessment, ACLD (cat. no. 2080.5). The ABS publishes extensive information on Census Data Quality.

ACCESSIBILITY

The Australian Census Longitudinal Datasets, 2006-11, 2011-16 and 2006-11-16 can be accessed through TableBuilder and the DataLab.

These microdata products are available to approved users. Users wishing to access the microdata should read the How to apply for Microdata web page, before applying for access by emailing microdata.access@abs.gov.au. Users should also familiarise themselves with information available via the Microdata Entry Page.

Any questions regarding access to microdata can be forwarded to microdata.access@abs.gov.au or phone (02) 6252 7714.